A REVIEW OF ALTERNATIVE APPROACHES TO INVENTORY THEORY

G. Hadley and T. M. Whitin

PREPARED FOR:
UNITED STATES AIR FORCE PROJECT RAND
A REVIEW OF ALTERNATIVE APPROACHES TO INVENTORY THEORY

G. Hadley and T. M. Whitin

This research is sponsored by the United States Air Force under Project RAND—Contract No. AF 49(638)-700 monitored by the Directorate of Development Plans, Deputy Chief of Staff, Research and Development, Hq USAF. Views or conclusions contained in this Memorandum should not be interpreted as representing the official opinion or policy of the United States Air Force.

DDC AVAILABILITY NOTICE
Qualified requesters may obtain copies of this report from the Defense Documentation Center (DDC).
In recent years both the military and industry have shown increasing interest in the use of analytic techniques as aids to inventory management. The following are some of the reasons: (1) inventory investment has increased steadily in both industry and the military; (2) new operational requirements in the military, and competitive conditions in industry, have intensified the pressure for faster and more reliable service from the inventory system; (3) industry constantly strives to maintain or improve profit margins, and the military to use budgets more efficiently; and (4) high-speed digital computers have made it possible to implement more sophisticated control procedures. Consequently, more people now than formerly are attempting to apply existing theory to practical problems, or are working on extensions of theory.

It seems to be an opportune time for a study which surveys the general field of inventory theory, summarizes past work, and highlights some of the important unsolved problems. This report is an attempt at such a survey. It discusses the usefulness and limitations of the theory in practical applications, but does not endeavor to survey inventory management techniques as they are actually being practiced in business and the military.

The study also excludes a number of topics which possibly could be grouped under the heading of inventory theory, such as the relationships between inventory fluctuations and business cycles, the control of inventory losses and shortages (due to theft, breakage, mislocation, etc.), and the problems of warehouse layout.
Instead, the study deals with that subset of problems centering on the fundamental questions of when to order and how much to order, for a system consisting of one or more stockage points. As the term is used here, then, inventory theory denotes that portion of the subject usually dealt with in operations research or management science.

This review is principally intended for readers who are reasonably familiar with probability theory and the calculus. Except for the greater part of Chap. 2, however, the study should be suitable for readers who wish an introduction to the subject but are not interested in the detailed mathematics.

G. Hadley is currently Professor of Economics at the University of New South Wales, Australia; T. M. Whitin is Professor or Economics, Wesleyan University. Both authors are consultants to the Logistics Department, The RAND Corporation.
SUMMARY

This study surveys the current state of inventory theory, and, so far as practical application is concerned, examines the usefulness and limitations of existing theoretical results. The report does not survey the actual techniques of inventory management employed by commercial or military organizations. We restrict our attention to inventory theory in the generally accepted meaning of the term: i.e., the analysis of problems associated with the control of stocks of physical goods at one or more locations and under various conditions.

The "Introduction" reviews the structure of inventory systems in terms of spatial configuration of the stocking points, the type of control methods used, the characteristics of the items themselves, and the nature of the stochastic processes. Then follows a brief outline of the methods that have been used to determine operating doctrines for inventory systems. Chapter 2, "Review of Analytical Models," forms the core of the study. There we consider the basic questions of when and how much to order under a variety of assumptions for any particular item. We consider both steady-state models (the stochastic processes remaining stationary in time) and dynamic models (where the stochastic processes change in time). Most of the models assume that if more than a single stocking point exists there are no interactions between the stocking points. Chapter 3 examines the various ways in which simulation can be used as a tool for studying inventory systems, and Chap. 4 explains what is here referred to as the Heuristic-Intuitive approach. Important unsolved problems are discussed in Chap. 5, and, finally, Chap. 6 examines difficulties of practical application. A list of references is appended.
Inventory system management has been, and still is, largely based upon intuitive judgment and experience. Theoretical systems, however, have received more and more attention, especially in the last ten years. Much of this current research has involved the application of mathematical analysis to the development of optimal inventory policy. Some investigations have applied simulation to these problems. We have attempted in this study to survey existing work on inventory theory, to enumerate certain problems that require investigation, and to consider some major obstacles in applying the theory to real-life systems.

The fundamental decisions that must be made in the operation of any inventory system are when to order and how much to order. Traditionally, inventory theory has been concerned with applying mathematical analysis as a tool to aid the inventory manager in making these decisions. This is done by constructing a mathematical model of the system. Normally, part of this model will be an expression which represents the cost of operating the system as a function of the variables appearing in the model. To obtain the best of all the policies for deciding when to order and how much to order, one minimizes this cost function. It is true, of course, that various types of mathematical models are needed to encompass the different situations encountered in practice.

Our survey of analytical techniques will begin with the simplest inventory models, i.e., those for which the demand rate and procurement leadtime are assumed to be known with certainty, and do not vary with time. Historically, these were the first types of inventory models
developed. Although it is realistic in certain cases to assume that the demand rate is known with certainty, it is more often true that the demand must be described probabilistically. The next group of models analyzed is concerned with making this generalization in situations where the stochastic processes do not change with time, i.e., to steady-state problems. Included are both periodic review models in which procurement decisions are made only at equally spaced intervals in time (perhaps every three months, for example), and lot-size reorder-point models in which an order is placed whenever the inventory reaches a level called the reorder point. For some models, the procurement lead-time must be a constant, while for others it can be treated also as a random variable. The review of steady-state models is concluded by showing how the techniques of dynamic programming can be used to prove the optimality of certain policies.

Dynamic situations in which the mean rate of demand can change with time are generally much more difficult to treat than the corresponding steady-state problems. Various dynamic models have been developed. We survey these, beginning, as we did for the steady-state models, with those which assume that the demand is known with certainty. Then we move to those that treat demand as a stochastic variable. Usually, dynamic programming furnishes the mathematical approach used both for the theoretical analyses and for obtaining numerical solutions. Only periodic review models are considered, since no dynamic versions of reorder-point type models have been developed. Furthermore, a finite planning horizon is used with dynamic models, i.e., only a finite number of periods is considered. Often, dynamic models deal with the combined
problem of production scheduling and inventory control, under the assumption either that production is carried out continuously or carried out in lots. This tends to increase the complexity of the models even more, because now another variable (or several variables) associated with the size of the workforce in each period appears. Because numerical computation using dynamic programming is often extremely time consuming, even when the largest digital computers are used, simplified but useful versions of some dynamic models have been obtained that require considerably less computational effort. These models are also reviewed.

All of the models referred to above ignore interactions between various stocking points and treat them as independent. The remainder of Chapter 2 surveys various theoretical results that have been obtained for multi-echelon systems. A new feature arises in the study of the dynamic behavior of such systems -- as a whole they may be dynamically unstable or may amplify periodic inputs. Although relatively little theoretical work has been done in analyzing the dynamic behavior of inventory systems, much has been done in electrical engineering. It seems that an abundance of this material is applicable, at least in part, to the analysis of inventory systems, and we survey those parts that seem most relevant.

When a mathematical model of an inventory system becomes too complicated to study analytically, one will often use simulation to analyze the system. Chapter 3 summarizes the various ways simulation is useful. Normally, simulation is not used to determine an optimal operating doctrine. Instead, simulations are used to try out specific operating doctrines and to obtain heuristic insights into the structure and behavior of inventory systems.
Although inventory theory has proved useful in analyzing complex problems, many additional issues remain to be studied. Many real-world systems deal with the complementary items that may be related to each other in use, or in terms of overall procurement budgets, or in the sharing of such common facilities as warehouses, transportation, and administration. Moreover, many important inventory systems, particularly the military variety, are of the multi-echelon type, consisting of a number of stocking points, some of which serve as warehouses for others. Except in very special circumstances, optimal inventory policies for determining where to stock and how much to stock at these multiple locations have not been obtained. Managers are also confronted by the problem of redistributing stock between the several points in a system; here, too, no definitive theory has been developed. The control of inventory levels through budget allocations, the use of aggregate indices, and other similar techniques have received relatively little detailed theoretical analysis, although many real-world systems might use such over-all control mechanisms. Limited research has been devoted to the interactions between the repair and procurement decisions for repairable items. Further development of inventory theory to take account of changing availabilities of procurement and repair resources, of varying demands, and of changing stocks at the several locations would also help to meet practical needs.

Even though it may be assumed that inventory theory will continue to advance and suggest optimal policies for more varied situations, difficulties will still persist in the practical application of such theory. The biggest problem often lies in the unavailability of data needed to estimate the parameters of the inventory models and to predict
future demand. In addition, it is frequently difficult to obtain personnel qualified to operate the system. Finally, incompatibilities between the characteristics of real-world systems and inventory theory make it necessary to adapt this theory or to recognize its limitations. This points to a continuing need for evaluation of the models used; we must adjust them to unfolding experience and make changes as circumstances and knowledge permit. This is a very challenging and difficult task.
FOREWORD

During a consultancy at The RAND Corporation in 1961, George Hadley and Tom Whitin were asked to offer a seminar in inventory theory. They have extended their original seminar paper into the present Memorandum. As prominent theoreticians in this field, the authors' summarizations and appraisals of their selected areas are valuable and noteworthy.

Murray A. Geisler, Head
Logistics Department
The RAND Corporation
CONTENTS

PREFACE .......................................................................................... iii

SUMMARY ......................................................................................... v

FOREWORD ..................................................................................... xi

Chapter

1. INTRODUCTION ................................................................. 1
   Historical Foundations ...................................................... 1
   The Structure of Inventory Systems ............................... 4
   Spatial Configurations .................................................. 4
   Type of Control .............................................................. 6
   Kind of Items Carried ..................................................... 7
   Processes Generating Stochastic Variables ................. 8
   Methods of Determining Operating Policies for Inventory
   Systems ................................................................. 9

2. REVIEW OF ANALYTICAL MODELS ................................. 13
   Classification of Analytical Models ......................... 13
   Deterministic Models with a Constant Rate of Demand .. 14
   Lot-Size, Reorder-Point Models with Stochastic
   Demands ............................................................ 26
   Variations of the (Q,s) Models ................................. 40
   Periodic Review Models ........................................... 42
   Alternative Stockout Procedures ............................. 53
   Use of Functional Equations to Study Nature of
   Optimal Operating Procedures ............................... 54
   Summary of Mathematical Techniques Used in the Study
   of Steady State Stochastic Models ......................... 63
   Static One-Period Models ........................................... 66
   Dynamic Models ....................................................... 69
   Deterministic Dynamic Models ............................... 70
   Stochastic Dynamic Models .................................... 77
   The Prediction Problem ............................................ 88
   Redistribution Models .............................................. 94
   Multi-Echelon Models ............................................. 95
   Dynamic Stability and Frequency Response of Complex
   Systems - Deterministic Inputs ............................... 97
   Dynamic Behavior - Stochastic Inputs ................... 117

3. SIMULATION ................................................................. 125
   Various Uses for Simulation .................................... 125
   Simulation and the Dynamic Behavior of Systems ...... 129

4. HEURISTIC-INTUITIVE APPROACH - ART OR SCIENCE? ...... 141

5. IMPORTANT UNSOLVED THEORETICAL PROBLEMS ............ 143
   Multi-Outlet and Multi-Echelon Installations ............. 143
   Product Interactions ............................................... 144
   Multi-Stage Production ........................................... 145
Interactions between System Behavior and Demand .......................... 146
Redistribution Aspects .................................................................. 147
Stochastic Leadtimes .................................................................... 148
Periodic Review Versus Transactions Recording ......................... 150
Repair Models .............................................................................. 151
Aggregate Versus Individual Item Control ................................... 152
Budget Constraints ....................................................................... 152
Lost Sales Models ........................................................................ 153

6. DIFFICULTIES OF PRACTICAL APPLICATION .............................. 154
   The Problems ............................................................................ 154
   Concluding Note ........................................................................ 161

REFERENCES .................................................................................. 163
Chapter 1

INTRODUCTION

HISTORICAL FOUNDATIONS

Inventory theory is the theory of storage. It is often impossible to meet the demands for some item unless it is stored in one way or another, as dams store water to be used in dry seasons, and retail stores keep stocks of merchandise to meet customer demands. Inventory problems are ancient, but have received detailed analysis only since the turn of the century -- mostly in the last fifteen years. This report surveys the present state of inventory theory, its relevance for practical applications, and some of the important unsolved problems.

Although work has been done on such subjects as the theory of dams, (1)* most inventory research has dealt with the stockage problems of military, industrial, and retail establishments. The present study also focuses on these problems, which make up the field of inventory control as it is normally regarded by workers in the management sciences.

The initial impetus for the analysis of inventory problems seems to have arisen during the period 1910-1920 in manufacturing firms that produced a variety of items in lots, with fairly high setup costs for each production run. These firms sought to discover the optimal size of a run in order to minimize the combined costs of setup and carrying inventory; the attempt bore fruit in the simple and well-known lot-size formula. A number of investigators arrived at this result independently, the earliest derivation known to the authors being that of F. W. Harris of Westinghouse. (2) The first full-length book on inventory problems

---

* See esp. Chap. 11, and Ref. (9).
seems to be that by F. E. Raymond, whose *Quantity and Economy in Manufacture* appeared in 1951.\(^3\)

More recently, the new fields of operations research and management science have brought with them an emphasis on applying quantitative techniques to management problems, and have produced a substantial literature on inventory problems. Mathematicians, economists, and statisticians have thus become interested in theoretical aspects of inventory problems which originally were the concern only of business managers and engineers. (It is interesting to speculate, in passing, why economists were not the first to study these problems, since inventories can be important to the study of economic dynamics.)

The three military services have contributed a considerable amount of the work done on inventory problems, partly through the large research groups they support; the Army and Navy also contract for research work by outside groups. It is only natural that the services should be highly interested in applying quantitative techniques to inventory management, since the sums they have invested in inventory are so huge that even a small percentage of saving could dwarf the outlay for the research program that made it possible.

A number of books dealing with inventory problems have appeared in recent years, with varying levels of mathematical sophistication and abstraction;\(^1, 4-13\) and almost all books on operations research devote at least one chapter to inventory problems. A great number of articles on inventory problems have been published, three of which are especially noteworthy for having laid the foundations for the abstract mathematical analysis of the subject. These are the articles by Arrow, Harris, and Marshak,\(^14\) and Dvoretzky, Kiefer, and Wolfowitz.\(^15\)\(^16\) The articles
that laid the foundation for the theory's practical applications are much more scattered.

Originally, the entire effort in this field was devoted to solving practical problems. Notable changes are taking place, however; one of the most important being the growing number of studies on the purely mathematical properties of abstract models which may have no immediate practical application at all. Both the advantages and the disadvantages of this shift are clearly recognized. A great advantage of abstract analysis is that it gives the subject a unity and generality which probably would be impossible otherwise and which, in turn, often suggest additional applications of the previously developed theory that might be overlooked. On the other hand, the widening emphasis on abstract analysis implies an increasing loss of contact with the practical problems that gave the theory's creation and growth their initial impetus. Research here is a two-way street; since practical problems often suggest new theoretical developments, neglecting them can lead to stagnation and sterility -- and people working in the field do tend to split into two groups, one working on theory and the other on application, and having little communication with each other. The same problems have afflicted mathematics and economics for some time.

It should be apparent, then, that we cannot use a single standard to evaluate all the work done in inventory control. It is improper to discredit a mathematical analysis of some hypothetical model because it is not immediately applicable to real-world problems, or because one of its assumptions does not conform with actuality. Nor is it proper to condemn proposed solutions of practical problems for not being
sufficiently rigorous, elegant, and general in their approach.

Some of the work to be reviewed here has no immediate practical application, but was developed instead to exhibit certain interesting mathematical properties of the system. Other models to be surveyed were developed explicitly for purposes of practical application. The reader should keep these differences in mind.

THE STRUCTURE OF INVENTORY SYSTEMS

In describing the structure of any inventory system, real or hypothetical, it is helpful to include:

(a) Its spatial configuration;

(b) The manner in which it is controlled (including an analysis of the flows of information, materials, funds, etc.);

(c) The kinds of items carried; and

(d) The processes generating the stochastic variables.

Spatial Configuration

The spatial configuration of an inventory system means its physical structure. For example, demands may occur at a single outlet or at a number of outlets. Outlets may be confined to a single city or scattered over several continents. An especially important distinction is that between single-echelon and multi-echelon systems.

In a single-echelon system all stockage points are on a par in the sense that no point serves as a warehouse for any other point. In a multi-echelon system one or more stockage points do serve as warehouses. The number of echelons is determined by the number of levels at which warehouses also act as stockage points for other warehouses. Figure 1 depicts a typical two-echelon system, a source-warehouse-retailer
structure (the source is not usually considered an echelon). How materials and information flow in such a system depends on how it is controlled.

![Diagram of a two-echelon inventory system]

Fig. 1 - A two-echelon inventory system

Another important question is whether the system has one, two, or several sources of supply for a given item; and a similar question is whether we must consider the inventory system as part of a production-distribution system coupled to the factory (or factories) through production-scheduling rules, or whether it can be studied independently of a production system.

A final possibility is the multi-stage system, in which a number of stockage points store different items. The stocking points are not independent, but are coupled together in some way. A typical multi-stage example is a production line having a number of points along the route to store the semi-finished product and/or raw materials.
Type of Control

An analysis of the way the system is controlled should include such things as whether control is centralized or decentralized, when decisions are made, what the rules are for making decisions, what is known about the system at any point in time, and how the flows of materials, information, and funds can occur.

An important aspect of any inventory system is what is known about its condition at any point in time. Two possibilities are of interest. The first, called transactions recording, determines the state of the system after each transaction (demand, receipt of items, ordering, etc.). Transactions recording makes decisions possible after each demand (whether or not to order, for example). Whether such a system is practical depends on what sort of information is required to make the decision. In some situations, nothing more is necessary than a card in the stock bin indicating it is time to reorder when the stock gets down to the level of the card. Other situations may require the use of a large-scale data-processing system where all transactions are fed into a digital computer that keeps and updates all records, prints out order requisitions, etc. The second possibility, called periodic review, determines the state of the system only at discrete, usually equally-spaced intervals in time. Decisions occur only at these review times (except perhaps when some emergency arises).

For multi-outlet and/or multi-echelon systems the extent to which decision-making is centralized or decentralized has an important bearing on the system analysis. Similarly, it is important to know what information the manager has available at the time he makes decisions.
System operation is also influenced by whether or not redistribution is allowed between various stocking points at a given echelon, and if so, by what modes of transportation. For multi-echelon structures, it is necessary to know whether any warehouse can ship to any stocking point at the next lowest level or whether it ships only to certain ones. Finally, it is necessary to know whether one or several different echelons procure directly from the source.

For any inventory system, the manager must know the nature of budgetary restrictions, constraints on available warehouse floorspace, or any other relevant forces. He often needs to know the nature of the contractual arrangements, if any, between the source and the inventory system (such as open-end or one-time negotiation).

Kinds of Items Carried

Some inventory systems handle only a single item while others carry as many as 1,500,000. These items may be substitutes for one another; they may be reparable or non-reparable (this is of consequence only if the reparables reappear in the inventory system again to be repaired); they may be perishable or non-perishable; they may become obsolete very rapidly or they may be staple items. All these characteristics have an important influence on the systems' behavior.

A final item characteristic is the nature of the costs incurred by the inventory system carrying it. First is the item's variable cost, which may or may not depend in an important way on the quantity procured. Then there are the transportation costs from the source, crating and uncrating costs, inspection costs, etc. There are also the administrative costs involved in placing orders and perhaps setup
costs at the factory. Various expenses are also associated with holding the item in inventory. These include insurance, cost of floorspace, rate of return requirements, etc. Finally, there are the costs incurred if demands occur when the system is out of stock. These costs are very difficult if not impossible to measure in many cases; however, they must be taken into account implicitly, if not explicitly, when one analyzes an inventory system.

Processes Generating Stochastic Variables

It is almost never true that the time pattern of demands impinging upon an inventory system can be predicted with certainty. Sometimes, however, there is sufficient regularity in the demand so that it can profitably be treated as deterministic. In most cases, though, the demand must be described in probabilistic terms.

For a single installation, the first important question is whether there is a single process or several different ones generating demands. We can characterize any given process which generates demands by specifying the probability distribution for the times between demands and the distribution describing the random variable for the number of units needed per demand. Important variations in the system operation arise depending on whether the mean rate of demand remains constant over time, whether the demands are correlated in time, and on the absolute level of the mean rate of demand. When demands can occur at more than one outlet, it is necessary to know how the demands are related -- i.e., whether or not they can be treated independently.

Given that a certain process generates demands on the system, the inventory system may or may not know what that process is. In the real
world, the system will almost never know, a priori, what sort of process generates demands. Information about the nature of this process is usually gained only from historical data. How well an inventory system operates depends to a considerable extent on how much its managers learn about the demand generation process.

Arising from the nature of the demands, a final important difference among inventory systems is what happens if a demand occurs when the system is out of stock. Basically, there are two situations. First, demands arising when the system is out of stock may be lost forever (i.e., the customer goes elsewhere). This is the lost-sales case. Second, demands occurring when the system is out of stock may be backordered and met when a procurement arrives. This is the backorder case. In some situations, it is also possible to have some sales lost and some backordered when the system is out of stock, depending perhaps on the backorders existing when a demand occurs.

METHODS OF DETERMINING OPERATING POLICIES FOR INVENTORY SYSTEMS

Three different procedures are available as aids in developing a set of operating rules for any inventory system. These may be referred to as analytical, simulation, and heuristic-intuitive techniques. The analytical approach consists of constructing a mathematical model of the system to be studied, and then using one of two procedures. The first is to determine the set of operating rules that optimize the model's operation (usually this means minimizing expected costs). The second is to select a specific operating rule that involves one or more parameters and to optimize with respect to these parameters. The analytical model will usually be an abstraction of the real world and will require
the introduction of a number of simplifying assumptions in order to handle it analytically. Thus, even though one finds a set of operating rules that optimizes the model, it does not follow that when the rules are applied to the real-world system its operation will be correspondingly optimized. It is desired, however, that these rules will form an efficient set, such that the incremental cost of improving upon them will not be offset by the expected additional savings.

The development of digital computers has opened new avenues for the study of inventory problems. For example, high-speed digital computers obtain numerical solutions to much more complex analytical models than would otherwise be possible. Digital computers also make feasible the use of simulation for studying inventory problems. To perform a simulation, one begins with a model of the system and a complete set of rules for operating the system. Then the computer generates the inputs (stochastic or deterministic). These include times between demands, number of units needed per demand, and perhaps leadtimes. Input generation proceeds in such a way that it simulates the actual processes that are assumed to be generating demands, leadtimes, etc. The computer uses the inputs to simulate the operation of the system through time, and determines the inventory levels, backorders, and all other quantities of interest. It determines when to place an order; places the order at the appropriate time; and if the procurement leadtime is not determined by other parts of the system, generates the leadtime for that particular order. It is possible in this way to simulate many months or years of system operation.

Simulation techniques have been useful in comparing some set of operating rules (obtained by optimizing an analytical model) with the
existing procedure for operating some real-world system. Instead of immediately trying out the new rules on the real system, the new rules and the existing rules can be simulated on a computer to see whether or not the new ones seem to lead to any substantial improvement in the system's operation. Sometimes, instead of having the computer generate random numbers to represent times between demands, etc., actual historical data may be used.

Simulation has been attempted in studying very complex inventory systems that are difficult or impossible to examine analytically. Simulation is also used to examine various sorts of operating rules. Those to be tested are usually obtained from heuristic considerations. The computer can carry out such simulations when given a complete set of rules for operating the system, or the computer can generate the random numbers and do the computations, while the personnel make the important operating decisions when needed. These are man-machine simulations. They add a further element of reality because the operators making the decisions can make mistakes, interpret the operating rules incorrectly, etc.

A final use of simulation is as a substitute for analysis when the analytical work becomes especially difficult to carry out. One particular area of this sort concerns the study of the stability and dynamic response of systems.

We have discussed the use of analytical models to optimize operating rules with respect to a given set of parameters, and the use of simulation as an aid in selecting operating rules. There remains the
question, where do the ideas for the operating rules originate?
Almost invariably, they arise from heuristic-intuitive considerations and are based on the formulator's experience with and intuitive feelings concerning the system under consideration. In addition to providing the inputs for analytical models and simulations, heuristic-intuitive considerations alone have supplied the operating rules for almost all real-world inventory systems, especially the more complex ones.

The following chapters will survey the analytical models which have been developed, the areas of application for simulation, and the heuristic-intuitive method. We will attempt to indicate some of the important unsolved problems. Finally, we will discuss the difficulties encountered in applying the theoretical models to practical problems.
Chapter 2

REVIEW OF THE ANALYTICAL MODELS

CLASSIFICATION OF ANALYTICAL MODELS

Only a very small subset of the various kinds of inventory systems referred to in "The Structure of Inventory Systems" have been studied analytically. In fact, when attempts are made to optimize the system, almost all effort has been restricted to studying systems consisting of a single stage and a single outlet. Relatively little has been done with multi-echelon, multi-stage, or multi-outlet models. Later we will indicate what has been done in these areas with regard to optimization models. The single installation models can be conveniently classified, for a particular item, according to the following categories:

Stochastic Processes

A. Demand:
   1. Deterministic or stochastic
   2. Constant or varying mean rate of demand
   3. Backorders or lost sales
   4. Units demanded one at a time or order size is a random variable
   5. Markovian or non-Markovian

B. Lead Times:
   1. Constant or stochastic
   2. Independent of or dependent upon each other

Control

Transactions recording or periodic review

Costs

A. Ordering cost fixed or not fixed

B. Unit cost of item a linear or nonlinear function of quantity ordered
C. Cost of backorder linear or nonlinear in number of backorders incurred and in length of time for which backorder exists

Items

A. Can or cannot be stocked indefinitely
B. Interact or do not interact

DETERMINISTIC MODELS WITH A CONSTANT RATE OF DEMAND

As noted previously, it is rare that the demands on an inventory system are sufficiently regular to be considered strictly deterministic. They must instead be described in probabilistic terms. Still it is interesting to study deterministic inventory models in which the demand rate, together with the procurement leadtimes, is assumed to be known with certainty. These models are illuminating for several reasons:

(1) At times they can be used as aids in obtaining helpful approximate solutions to real-world problems;

(2) It is often desirable to compare results where uncertainty is involved with results where demand and other factors are assumed to be deterministic;

(3) The models often have an interesting structure in themselves.

Here we will focus attention only on situations in which the demand rate remains constant over time. First, we will study the simplest "lot-size reorder-point" model. This was the first analytical model developed, and as we suggest in the Introduction, F. W. Harris seems the first to have worked with it. The problem to be studied is that of controlling the inventory at a single installation when the demand rate is a constant $\lambda$ units per year independent of time. The procurement leadtime $\tau$ is assumed to be a constant, independent of all other
variables or parameters. Furthermore, we will assume that when an order is placed, the entire quantity ordered is delivered as a single package -- i.e., it never happens that an order is split and that part of it arrives at one time and part at another time. The basic problem is determining how much to order and when to place an order. The procedure for computing these variables employs the minimization of the long-run average annual variable costs.* This is equivalent to maximizing average yearly profits if the selling price is constant, since the yearly revenues will be $\lambda p$ where $p$ is the unit price.

To find the average annual variable cost, it is necessary to include only those costs that are functions of the variables to be determined. For our model we will assume that there are only two relevant cost terms: the cost of ordering and the cost of holding inventory. Imagine that each time an order is placed, no matter what its size, a fixed cost $A$ is incurred. An additional cost, incurred for storage and proportional to the time a unit is held in stock, is $IC$ per year, where $0 < I \leq 1$ and $C$ is the unit cost. The unit cost of the item is assumed to be constant, independent of the quantity ordered. For this reason, the average annual cost of the units is $\lambda C$ and is independent of the order quantity or the reorder rule, thereby implying that these costs may be excluded from the cost expression to be minimized. It will also be required that the system always be stocked when a demand occurs (such a restriction can be legitimately made in the deterministic case). Thus, there will never be any stockout costs.

*One might instead determine the decision variable by minimizing the discounted value of all future costs. For inventory problems, the use of either criterion will lead to essentially the same values of the decision variables.
Because there is a cost incurred each time an order is placed, it will generally be desirable to incur some inventory carrying costs instead of placing an order each time a demand occurs. We will assume that the system's operating doctrine is to place an order for $Q$ units after every $Q$ demands. $Q$ is a variable to be determined. This operating doctrine is a reasonable one since, with no uncertainty, and no changes over time, if it is ever optimal to place an order for $Q$ units, the same quantity will be ordered each time an order is placed. In determining $Q$, it will be imagined that $Q$ and the quantity demanded in any time period can be treated as continuous variables.

It is clear that if inventory carrying costs are to be minimized, the onhand inventory should reach zero just as an order arrives. If a positive quantity were on hand, unnecessary carrying charges would be incurred. It cannot reach zero before the order arrives, since this would imply that demands occur when the system is out of stock -- which has been prohibited above. This condition shows that the onhand inventory will vary between $Q$ and 0, being $Q$ immediately after the arrival of an order, and 0 immediately before the arrival. This is illustrated in Fig. 2. The system is said to go through a cycle of operation in the time between the placement or receipt of two successive orders. Since an order is placed after every $Q$ demands, the length of a cycle is $T = Q/\lambda$ years.

The observation that the onhand inventory reaches zero just as an order arrives makes it possible to develop a rule for reordering. It is convenient to express this rule in terms of a reorder point $s_h$, i.e., when the inventory level reaches $s_h$ an order is placed. Let $m$ be the largest integer less then or equal to $\tau/T$ ($\tau$ being the procurement leadtime). Then it can be
seen in Fig. 2 that for any $Q$,

\[ s_h = \mu - mQ \]

where $\mu = \lambda \tau$ is the leadtime demand, i.e., the amount demanded in the procurement leadtime. Note that $mQ$ is the amount on order when the reorder point is reached. Often, instead of dealing with the onhand inventory level, it is convenient to deal with the amount onhand plus that on order. The reorder point $s$, in terms of the onhand plus on order inventory level, must then be $\mu$, i.e., $s = \mu$.

![Diagram showing onhand inventory over time with reorder point](image)

**Fig. 2**—Time behavior of onhand inventory for simplest lot-size model

It remains to set up the cost equation and to determine the optimal ordering quantity, denoted by $Q^*$. Since the length of a cycle is $Q/\lambda$, the long-run average number of orders placed per year will be $\lambda/Q$ and the corresponding average annual fixed cost of placing orders will be $\lambda A/Q$. If $x$ units are in inventory at time $t$, the inventory carrying charges incurred between $t$ and $t + dt$ will be $IC x dt$. At the beginning of a cycle, the onhand inventory is $Q$ and at a time
t after the cycle begins, the onhand inventory is \( Q - \lambda t \). Thus the inventory carrying charges per cycle are

\[
(2) \quad \int_{0}^{T} (Q - \lambda t) \, dt = IC \frac{Q^2}{\lambda},
\]

since \( T = Q/\lambda \). The average inventory carrying cost per year is then the cost per cycle times the average number of cycles per year. This yields \( ICQ/2 \).

Since (see Fig. 2) the average inventory is \( Q/2 \), it is clear from the above observation that the average annual cost of holding inventory is \( ICQ/2 \).

The average annual cost of ordering and holding inventory \( R \) is then

\[
(3) \quad R = \frac{\lambda}{Q} A + \frac{IC}{2} Q.
\]

Note that \( R \) is differentiable for all \( Q > 0 \), and \( R = \infty \) at \( Q = 0 \) or \( \infty \), while \( R \) is finite everywhere else. Thus \( Q^* \) which yields the absolute minimum of \( Q \) must satisfy

\[
(4) \quad \frac{dR}{dQ} = 0 = -\frac{\lambda}{Q^2} A + \frac{IC}{2},
\]

or

\[
(5) \quad Q^* = \sqrt{\frac{2\lambda A}{IC}} = Q^*_{\lambda},
\]

since Eq. 4 has the unique solution Eq. 5. Equation 5 gives the optimal order quantity in terms of the other system parameters. Note that, other things being equal, the optimal order quantity (and the average inventory \( Q^*/2 \)) increases proportionally with the square root of the sales rate and in inverse proportion with the square root of the cost.

All this points out that different items need individual treatment. For instance, if a number of items are stocked, it is not generally optimal to have \( Q^* \) equal to the same number of weeks of supply for each
item. The \( Q^* \) of Eq. 5 is often called the economic lot size or the Wilson \( Q \). The symbol \( Q_w \) represents the square root in Eq. 5.

If Eq. 5 is substituted into Eq. 3, then \( R^* \), the minimum average annual cost of ordering and holding inventory is:

\[
R^* = \sqrt{2 \lambda A C} = R_w
\]

The symbol \( R_w \) will be used for the square root in Eq. 6. When a \( Q \) value other than \( Q^* \) is used as the ordering quantity, then the ratio \( R/R^* \) can easily be found by dividing Eq. 3 by Eq. 6 and is:

\[
\frac{R}{R^*} = \frac{1}{2} \left[ \frac{Q}{Q^*} + \frac{Q^*}{Q} \right]
\]

Here we have an equation relating \( R/R^* \) to \( Q/Q^* \) which is completely independent of the system parameters. It shows that when \( Q \) differs from \( Q^* \) by a factor of 2, \( R \) will be only 25 per cent greater than \( R^* \).

Consider next a slightly different situation in which the inventory system is a factory warehouse. Suppose that the item is produced in lots and that the production rate is \( \psi \) (\( \psi \) is assumed to be greater than \( \lambda \), the rate of demand, for the units at the warehouse), and that units flow into the warehouse at a constant rate of \( \psi \) while a lot is being produced. The net rate of influx during production is then \( \psi - \lambda \); with no production the net efflux is \( \lambda \). If the lot size is \( Q \), then the onhand inventory at the warehouse will behave as in Fig. 3. Again, suppose that requirements call for a stocked warehouse whenever a demand occurs. The time required to produce the lot will be \( T_1 = Q/\psi \).
Fig. 3 — Behavior of onhand inventory in a factory warehouse

The maximum inventory in the factory warehouse is reached just as production of the lot is finished and is

\[ T_1 (\psi - \lambda) = Q \left( 1 - \frac{\lambda}{\psi} \right) . \]  \hspace{1cm} (8)

The time required to deplete the inventory is then

\[ T_2 = \frac{Q}{\lambda} \left( 1 - \frac{\lambda}{\psi} \right) , \]  \hspace{1cm} (9)

and \( T \), the length of the cycle, is \( T_1 + T_2 = Q/\lambda \).

If \( A \) is the setup cost for a production run and \( IC \) is the cost per unit year of holding a unit in inventory, then the average annual cost of setups and holding inventory is

\[ R = \frac{A}{Q} + IC \frac{Q}{2} \left( 1 - \frac{\lambda}{\psi} \right) . \]  \hspace{1cm} (10)

Differentiating with respect to \( Q \), we find that the optimal \( Q \) is

\[ Q^* = Q \sqrt{\frac{\psi}{\psi - \lambda}} \left( \frac{1}{2} \right) . \]  \hspace{1cm} (11)
$Q^* > Q_w$ since the entire lot is never in inventory at one time.

Let $m$ be the greatest integer less than or equal to $\tau/T$, where $\tau$ is the time elapsed between placement of a production order and the time the first unit comes off the line. Then if $\tau - mT < T_2$, the reorder point based on the onhand inventory level is $s_h = \mu - mQ$ as in the simple lot size case. However, if $\tau - mT > T_2$, the reorder point is

\begin{equation}
(12) \quad s_h = \mu - \eta + (m+1) \left( \frac{\lambda}{\lambda} - 1 \right) Q,
\end{equation}

where $\eta = \frac{\lambda}{\lambda}$.

Consider once again the case where the entire order arrives in a single batch. In the above we assumed that the system would never be out of stock when a demand occurred. We will now assume that the system is allowed to be out of stock and that demands occurring at such a time are backordered. A cost, however, will be associated with each backorder. The variety of forms which the cost might take is infinite. In general, it will depend upon how long the backorder exists. Here we will assume that each backorder costs $\Pi + \hat{\Pi} t$, where $t$ represents the length of time the backorder exists. The cost is composed of a fixed cost plus another charge that is proportional to $t$. The average annual cost of backorders is then $\Pi$ times the number of backorders incurred per year plus $\hat{\Pi}$ times the unit years of shortage, where the latter is found by integrating the backorders at any point in time over time.

Consider then a situation where backorders are allowed. Let $v \geq 0$ be the number of backorders on the books when an order arrives.
Figure 4 illustrates the situation. It is assumed that when an order arrives, all outstanding backorders are filled if this is possible.

![Diagram](image)

Fig. 4 — A system in which backorders are allowed

In this situation, two variables must be determined, $Q$ and $v$.

Observe that

$$(13) \quad T_1 = \frac{Q - v}{\lambda}; \quad T_2 = T - T_1 = \frac{v}{\lambda}.$$ 

Thus the average annual cost of backorders is

$$\frac{1}{Q} \left[ \pi \lambda v + \frac{1}{2} \pi v^2 \right].$$

The inventory carrying cost per cycle is

$$\text{IC} \int_{0}^{T_1} (Q - v - \lambda t) \, dt = \frac{\text{IC} (Q - v)^2}{2\lambda};$$

hence the average annual inventory carrying cost is $\text{IC} (Q - v)^2/2Q$. 
The average annual variable cost of ordering, holding inventory
and incurring backorders \( R \), is

\[
R = \frac{1}{Q} \lambda A + \frac{IC}{2Q} (Q - v)^2 + \frac{1}{Q} \left\{ \frac{1}{2} \mu v + \frac{1}{2} \frac{\hat{\pi}}{\pi} v^2 \right\}.
\]

It is desired to determine the values of \( Q, v \) which yield the absolute minimum of \( R \) in the region \( 0 < Q < \infty \), \( 0 \leq v < \infty \). Then if the optimal \( v \) satisfies \( 0 < v^* < \infty \), \( Q^* \) and \( v^* \) must satisfy the equations

\[
\frac{\partial R}{\partial Q} = 0 = -\frac{1}{Q^2} \left[ \lambda A + \frac{1}{2} IC (Q - v)^2 + \mu v + \frac{1}{2} \frac{\hat{\pi}}{\pi} v^2 \right] + \frac{IC}{Q} (Q - v)
\]

\[
\frac{\partial R}{\partial v} = 0 = -\frac{IC}{Q} (Q - v) + \frac{1}{Q} \lambda A + \frac{1}{Q} \frac{\hat{\pi}}{\pi} v.
\]

It is not hard to show that if \( \frac{\hat{\pi}}{\pi} = 0 \), then either \( v^* = 0 \) or \( v^* = \infty \), and \( v^* = 0 \) if \( \lambda A > R_w \). When \( \lambda A = R_w \) any value of \( v \) is optimal.

Furthermore, even if \( \frac{\hat{\pi}}{\pi} \neq 0 \), \( v^* = 0 \) if \( \lambda A > R_w \). When \( \frac{\hat{\pi}}{\pi} \neq 0 \), \( v^* \) cannot be infinite. An infinite value of \( v^* \) means that backorders are accumulated forever and an order is never placed. In this case, the inventory system should not be operated at all. When \( \frac{\hat{\pi}}{\pi} \neq 0 \) and \( \lambda A < R_w \), the optimal values of \( Q, v \) are

\[
Q^* = \left[ \frac{\hat{\pi} + IC}{\frac{\hat{\pi}}{\pi}} \right]^{1/2} \left[ Q_w^2 - \frac{\lambda A^2}{IC(\frac{\hat{\pi}}{\pi} + IC)} \right]^{1/2}
\]

and

\[
v^* = \left[ \frac{\hat{\pi} + IC}{\frac{\hat{\pi}}{\pi}} \right]^{-1} \left\{ -\lambda A + \left[ r_w^2 (1 + \frac{IC}{\hat{\pi}}) - \frac{IC}{\hat{\pi}} (\lambda A)^2 \right]^{-1/2} \right\}.
\]
When \( \Pi = 0 \), these formulas reduce to

\[
Q^* = Q_w \left[ \frac{\hat{\eta} + IC}{\hat{\eta}} \right]^{1/2},
\]

\[
v^* = R_w \left[ \frac{\hat{\eta} (\hat{\eta} + IC)}{\hat{\eta}} \right]^{-1/2}.
\]

If \( \Pi = 0 \), it is always optimal to incur at least a very small quantity of backorders (in the continuous case) to avoid carrying units in inventory for the entire cycle to meet these demands at the end of the cycle.

A new inventory level is needed here to define a reorder point, since there might conceivably be no inventory on hand when an order is placed. Instead some backorders might be on the books. Let us define the net inventory as the quantity on hand minus the backorders. If there is inventory on hand, the net inventory is positive; if there are backorders, it is negative. The net inventory can be used to define a reorder point; it is \( s_n^* = \mu - mQ^* - v^* \) where as before \( m \) is the largest integer less than or equal to \( \gamma / \tau \). Sometimes it is convenient to use another inventory level, called the inventory position, which is the quantity on hand plus on order minus backorders. The optimal reorder point \( s^* \) in terms of the inventory position is \( s^* = \mu \).

Consider now the case where demands occurring when the system is out of stock are lost, instead of backordered. Let \( \tau \) be the cost of a lost sale; \( \tau \) includes the good will loss plus lost profit. Then it is very easy to see that if \( \lambda \tau \geq R_w \), it is never optimal to lose a sale, i.e., \( \hat{T} \), the time per cycle during which the system is out of stock
should be 0. On the other hand, if λτ < R_w, an order is never placed, i.e., the system should not be operated. One simple way to see that it is never optimal to lose any sales when λτ ≥ R_w is to note that if \( \hat{T} > 0 \), we can imagine rearranging the time sequence so that there is a long period with nothing but lost sales and another one with no lost sales. However, costs can be reduced by ordering in lots of size Q_w during the time when there is nothing but lost sales.

There are many other variations of the simple deterministic models considered above. For example, suppose that in the simplest lot-size model, it is necessary to order in multiples of a standard package size so that \( Q = nx \) where \( n = 1, 2, 3, \ldots \), and \( X \) is the number of units in the package. Now \( n \), rather than \( Q \) becomes the variable; \( n \) must be a positive integer, however. If demand is still treated as continuous then Eq. 3 becomes

\[
(21) \quad R(n) = \frac{\lambda A}{nX} + \frac{ICX}{2n}.
\]

The smallest \( n \) that minimizes \( R(n) \) must satisfy

\[
(22) \quad \Delta R(n) = R(n) - R(n-1) < 0; \quad \Delta R(n+1) = R(n+1) - R(n) \geq 0
\]

or \( n^* = 1 \).

Thus the optimal \( n \) is the largest \( n \) for which \( \Delta R(n) < 0 \) or is unity: i.e., the largest \( n \) for which

\[
(23) \quad n(n-1) - \frac{2\lambda A}{ICX^2}.
\]

It is also possible to take account of the integrality of demand and \( Q \). Suppose that units are demanded one at a time and that \( \hat{T} \) is the
time between demands, so that the average rate of demand $\lambda = 1/\hat{t}$. The system is required to have stock when a demand occurs. It is clear therefore that an order should arrive in the system precisely at time $\hat{t}$ after a demand has reduced the onhand inventory to zero, i.e., just as the next demand occurs. The maximum inventory is thus $Q - 1$ and the inventory carrying cost per cycle is

$$\text{IC} \sum_{i=1}^{Q-1} i = \frac{\text{IC}Q}{\lambda} \left( \frac{Q - 1}{2} \right).$$

Thus the average annual costs of ordering and holding inventory are

$$(24) \quad R(Q) = \frac{\lambda}{Q} A + \frac{1}{2} \text{IC}(Q - 1).$$

If $Q^*$ is the smallest $Q$ minimizing $R(Q)$, then $Q^*$ is the largest $Q$ for which

$$(25) \quad Q(Q - 1) < \frac{Q^2}{2}. \quad \text{We often refer to the models in this section as lot-size, reorder-point, or } (Q,s) \text{ types. The order size is fixed, and an order is placed each time the inventory level reaches the reorder point. The next section examines lot-size, reorder-point models with stochastic demands.}$$

**LOT-SIZE, REORDER-POINT MODELS WITH STOCHASTIC DEMANDS**

Two assumptions are necessary if the requirements of the lot-size, reorder-point model are to be fulfilled precisely when the quantity demanded in any time period is a stochastic variable. We must assume that a transactions recording system is being used, and that when demand is treated as discrete, units are requested one at a time. These
conditions must be met to avoid an "overshoot" of the reorder point. The \((Q,s)\) model supposes that an order is placed when the reorder point is reached and that there is no "overshoot." If the size of each demand constituted a random variable, then it would be impossible to guarantee no overshoot. In the real world, \((Q,s)\) models are often used even when the number of units per demand do constitute a random variable and where the system does not operate strictly using transactions recording. Then we assume that the overshoot of the reorder point is slight and can be ignored.

The change to stochastic demand introduces a problem not encountered in the deterministic situation. It was possible, with deterministic demands, to predict precisely what the onhand or net inventory would be when an order arrived. This is not true when the demand is described probabilistically. Hence it is usually desirable to carry some additional stock to avoid stockout costs that occur when the leadtime demand is greater than its mean value. The safety stock \(v\) will be defined in backorders cases as the expected value of the net inventory at the time an order arrives; it will be defined in lost sales cases as the expected value of the onhand inventory at the time an order arrives. Thus \(v\) is a measure of the inventory augmentation designed to avoid stockout costs which result because the leadtime demand is a random variable.

The very nature of a \((Q,s)\) model requires that the demand-generating process remain time invariant, particularly the mean rate of demand \(\lambda\). We will assume that the demand \(x\), in a time interval of length \(t\), can be described by a discrete or continuous density function.
To begin, we will present two simple approximate \((Q,s)\) models: one for the backorder case and one for the lost sales case. It is these simple models that are usually to be found in the literature\(^{(5, 8, 12)}\). Furthermore, when \((Q,s)\) models are applied in practice, it is almost always these same models that are used. The optimal values of \(Q\) and \(s\) are determined by minimizing the expected annual cost which, from the frequency definition of probabilities, must be equal to the long-run average annual cost.

The backorder model will be presented first. In its derivation, the following assumptions are usually made:

1. The item's unit cost \(C\) is a constant independent of \(Q\).
2. Each backorder costs \(\Pi\). There is no cost \(\Pi_t\) that depends on the length of time \(t\) the backorder exists, i.e., \(\Pi_t = 0\).
3. There is never more than a single order outstanding.

Because of this third assumption there are no orders outstanding at the time the reorder point is reached. The reorder point, \(s\), will be based on the onhand inventory level. This requires the following additional assumption:

4. When a procurement arrives, it is always sufficient to bring the onhand inventory level above the reorder point.

Assumption 4 is required when the reorder point is based on the onhand inventory level. Otherwise a very heavy demand in one cycle might create so many backorders that the arrival of the procurement would not bring the onhand inventory up to the reorder point, and hence another order would never be placed.
The cost expression will now be worked out under the assumption that \( Q, s \) and the demand variable may be treated as continuous. Let 
\[ f(x, t) \, dx \]
be the probability that the demand in a time interval of length \( t \) lies between \( x \) and \( x + dx \).

If \( \tau \) is the procurement leadtime, the expected number of backorders incurred per cycle is
\[
\int_{s}^{\infty} (x - s) f(x, \tau) \, dx;
\]

In the event that the procurement leadtime is also a random variable with density \( g(\tau) \), the expected number of backorders per cycle is
\[
\int_{0}^{\infty} \int_{s}^{\infty} (x - s) f(x, \tau) g(\tau) \, d\tau \, dx = \int_{s}^{\infty} (x - s) h(x) \, dx,
\]
where
\[
h(x) = \int_{0}^{\infty} f(x, \tau) g(\tau) \, d\tau
\]
is the marginal distribution of leadtime demand (if the leadtime is constant, then \( h(x) = f(x, \tau) \)). Since an order is placed after every \( Q \) demands there are on the average \( \lambda / Q \) cycles per year. Hence the expected annual cost of backorders is

(26)
\[
E(Q, s) = \frac{\lambda}{Q} \int_{s}^{\infty} (x - s) h(x) \, dx.
\]

The expected annual cost of carrying inventory, \( H(Q, s) \), is \( IC \) times the expected number of unit years of stock held per year, where \( IC \) is the cost of carrying one unit in stock for one year. To compute \( H(Q, s) \) we make another assumption:
(5) The expected number of backorders incurred per cycle is small, and incurred only toward the end of the cycle, so that the expected unit years of backorders acquired per cycle are negligible.

Recall that the net inventory is the amount on hand minus the backorders. The expected net inventory at any point then is the expected onhand inventory minus the expected backorders, or the expected onhand inventory is equal to the expected net inventory plus the expected backorders. However, when Assumption 5 holds, the integral of the expected onhand inventory over time is approximately equal to the integral of the net inventory.

Now the expected net inventory when a procurement arrives is the safety stock $v$. Immediately after the procurement arrives the expected net inventory is $Q + v$. Since the mean demand rate is constant, the expected net inventory will decline linearly from $Q + v$ at the beginning of a cycle to $v$ at the end and will average to $(Q/2) + v$, which is the expected unit years of stock held per year. Thus

$$H(Q, s) = IC \left[ \frac{Q}{2} + v \right].$$

It is easy to evaluate $v$ in terms of $s$. Simply observe that

$$v = \int_{0}^{\infty} (s-x)h(x)dx = s - \mu,$$

where $\mu$ is the expected leadtime demand. Consequently

$$H(Q, s) = IC \left[ \frac{Q}{2} + s - \mu \right].$$

The expected annual cost of ordering, holding inventory, and incurring
backorders is then

\[ R = \frac{\lambda}{Q} A + IC \left[ \frac{Q}{2} + s - s^* \right] + \frac{\Pi \lambda}{Q} \int_s^\infty (x-s)h(x)dx. \]  

If \( 0 < Q^* < \infty, \ 0 < s^* < \infty, \) then \( Q^*, \ s^* \) must satisfy the equations

\[ \frac{\partial R}{\partial Q} = 0 = -\frac{\lambda}{Q^2} \left[ A + \Pi \int_s^\infty (x-s)h(x)dx \right] + \frac{IC}{2}, \]

\[ \frac{\partial R}{\partial s} = 0 = IC - \frac{\Pi \lambda}{Q} H(s), \]

where \( H(x) \) is the complementary cumulative for \( h(x) \). Thus two equations are obtained and must be solved for \( Q, s \). They are:

\[ Q = \sqrt{\frac{2\lambda}{IC} [A + \Pi \eta(s)]}, \]

\[ H(s) = \frac{QIC}{\Pi \lambda}, \]

where

\[ \eta(s) = \int_s^\infty (x-s)h(x)dx. \]

is the expected number of backorders incurred per cycle. Note that the computations need only the marginal distribution of leadtime demand. One can use a distribution obtained from experience, if desired, in place of a theoretical distribution.

A useful iterative scheme for solving Eqs. 33 and 34 which always converges when a solution exists is to set \( Q = Q_1 \) in Eq. 34 and determine \( s \) -- call it \( s_1 \). Use \( s_1 \) in Eq. 33 to determine \( Q_2 \), then use \( Q_2 \) in Eq. 34 to determine \( s_2 \). Then use \( s_2 \) in Eq. 33 to determine \( Q_3 \), etc. Geometrically, Eqs. 33 and 34 can be thought of in the \((Q,s)\) planes as representing two curves which look like those in Fig. 5, which shows the way
in which the iterative scheme converges. The convergence is usually quite rapid. If \( h(x) > 0 \) for all \( x > 0 \), then \( R(Q,s) \) is a strictly convex function of \( Q \) and \( s \). Here the optimal solution is unique. Even if \( h(x) = 0 \) over some intervals of \( x \), \( R(Q,s) \) is still convex (but not necessarily strictly convex with respect to \( s \)) and it can here also be shown that \( Q^* \) and \( s^* \) are unique. If \( \Pi \) is small, Eqs. 33 and 34 may not have a solution.

\[
\begin{align*}
\text{Eq. 34} & & \text{Eq. 33} \\
\text{(Q*, s*)} & & \\
0 & & Q_w & & Q_2 & & \frac{\Pi \lambda}{IC}
\end{align*}
\]

**Fig. 5 — Graphical representation of iterative scheme**

In the event that \( Q, s \) are treated as discrete variables, the equations corresponding to Eqs. 33, 34 are

\[
(36) \quad Q(Q-1) < \frac{\partial \Pi}{IC} [A + \Pi \eta(s)]
\]

and
\[(37) \quad H(s) > \frac{QIC}{2}. \]

The iterative procedure is to begin with \( Q = Q_0 \) in Eq. 37 and to determine the largest \( s \), call it \( s_1 \), which satisfies the Eq. 37. Then \( s_1 \) is substituted in Eq. 36 and \( Q_2 \) is determined by finding the largest \( Q \) which satisfies Eq. 36, etc.

For the simple approximate models under consideration, the lost sales case differs little from the backorders. The expected number of lost sales per cycle is equal to the expected number of backorders per cycle, i.e., to \( \eta(s) \). If \( T \) is the average length of time for which lost sales are incurred per cycle, then the average number of cycles per year will be \( \lambda/(Q + \lambda T) \). For this simple model it is usually assumed that \( T \) is so small that the average number of cycles per year can again be expressed as \( \lambda/Q \).

In the lost sales case the safety stock \( v \) is the expected value of the onhand inventory at the time of arrival of a procurement. Thus

\[(38) \quad v = \int_0^\infty (s-x)h(x)dx = s - \mu + \eta(s). \]

The average annual costs of carrying inventory are still given by Eq. 27, so that

\[ H(Q, s) = IC \left[ \frac{Q}{2} + s - \mu + \eta(s) \right]. \]

The expected annual costs of ordering, holding inventory, and lost sales is

\[(39) \quad R = \frac{\lambda}{Q} A + IC \left[ \frac{Q}{2} + s - \mu \right] + \left[ \frac{\Pi \lambda}{Q} + IC \right] \eta(s). \]

Now, of course, \( \Pi \) is the cost of a lost sale not of a backorder. The
equations corresponding to Eqs. 33, 34 become

\[ Q = \sqrt{\frac{2A}{IC} \left[ A + \eta \beta(s) \right]} \quad H(s) = \frac{QIC}{\pi A + QIC} \]

If one computes \( Q^*, s^* \) from the lost sales model and the backorder model using the same set of parameters in each, it will turn out that \( s^* \) for the lost sales model is somewhat greater than \( s^* \) for the backorder model while \( Q^* \) for the lost sales model is somewhat less than \( Q^* \) for the backorder model. These differences usually will be extremely small. The two equations given in Eq. 40 always have a solution.

The two models developed above also apply in cases where more than a single order may be outstanding, provided that all of the assumptions except the one concerning the number of orders outstanding remain valid. The only change necessary in interpretation is to consider \( s \) to be the inventory position in the backorders case and to be the onhand plus on order inventory in the lost sales case.

The simple models above made a number of approximations. It is possible, however, to work out the exact equations for the backorder case with constant leadtimes and Poisson or normal demand distributions. The assumption that the demand is Poisson or normally distributed implies that the process generating demands is Markovian. The case of Poisson demands will be worked out here.

To begin, we will observe that the onhand or net inventory levels are not generally suitable for defining a reorder point, since if a particularly heavy demand occurred during one cycle, the arrival
of a procurement might not satisfy all the backorders. Hence the on
hand or net inventory might never get up to the reorder point again.
Instead, the "inventory position" must be used. Recall that the
inventory position is defined as the net inventory plus the amount on
order: i.e., the quantity on hand plus on order minus the backorders.
When the inventory position reaches the reorder point \( s \), then an order
for \( Q \) units is placed and the inventory position immediately goes to
\( s + Q \). Thus the inventory position can never be less than \( s \) or more
than \( s + Q \). It never remains in state \( s \) for a finite amount of time,
since an order is placed immediately after a demand moves the inventory
position from \( s + 1 \) to \( s \). Thus the inventory position of the system
at any point in time can have only one of the values \( s + 1, \ldots, s + Q \).
These will be referred to as "states" of the system. It is clear that
the inventory position can be used to define the reorder point and the
difficulties referred to in the models above will not occur.

Let \( \rho(j) \) be the steady state probability, which will be referred to
simply as the state probability, that the inventory position of the
system is \( s + j \) at any arbitrary point in time. By definition,
steady state implies that \( \rho(j) \) is independent of time. When the demand
is Poisson distributed, then during a time \( dt \) the probability that a
demand will occur is \( \lambda dt \); this is also the probability that if the
system is in state \( s+j+1 \) it will move to \( s+j \). If \( \rho(j) \) is to be inde-
dependent of time, then the probability that the system leaves state
\( j \) in time \( dt \) must be exactly equal to the probability that it move
from some other state to state \( j \). Thus, on dropping terms which
involve a higher power than the first in \( dt \),

\[
(41) \quad (\lambda dt)\rho(j+1) = (\lambda dt)\rho(j), \quad j = 1, \ldots, s + Q - 1; (\lambda dt)\rho(1) = (\lambda dt)\rho(Q).
\]
Consequently, \( \rho(j) = 1/Q \), \( j = 1, \ldots, Q \), since it must be true that the \( \rho(j) \) sum to unity. Hence the states for the inventory position are uniformly distributed.

It is now possible to compute the various expected values needed in setting up the cost expression. As before, let \( A \) be the cost of placing an order, \( IC \) the cost of carrying one unit for one year, and \( \Pi + \Pi \)t the cost of a backorder, where \( t \) is the length of time for which a backorder exists. The unit cost \( C \) of the item will be assumed to be constant, and hence it need not be included in the cost expression.

The expected annual cost \( R \) is

\[
R = \frac{A}{Q} + IC E(Q,s) + \Pi B(Q,s) + \Pi B(Q,s),
\]

where \( H(Q,s) \) is the expected onhand inventory at any point in time, \( E(Q,s) \) is the expected number of backorders incurred per year, and \( B(Q,s) \) is the expected number of backorders at any point in time.

From the definition of the inventory position it follows that \( H(Q,s) \) is equal to the expected value of the inventory position, minus the expected amount on order, plus \( B(Q,s) \). However, the expected amount on order is the expected leadtime demand \( \mu \) and the expected inventory position is

\[
\frac{1}{Q} \sum_{j=1}^{Q} (s+j) = \frac{Q}{2} + \frac{1}{2} + s ,
\]

so that

\[
H(Q,s) = \frac{Q}{2} + \frac{1}{2} + s - \mu + B(Q,s) .
\]
We may use a continuous pipeline to illustrate the principle that the expected amount on order is equal to the expected leadtime demand. Orders flow into the pipeline at the demand rate \( \lambda \). Each remains in the pipe for an average time \( \tau \). Hence the average quantity within the pipe is \( \lambda \tau \), which is the expected leadtime demand.

Now \( E(Q,s) \) is simply \( \lambda P_{\text{out}} \) where \( P_{\text{out}} \) is the probability that the system is out of stock at any time \( t \). To compute \( B(Q,s) \) and \( P_{\text{out}} \), note that if the inventory position of the system was \( s + j \) at time \( t - \tau \), the probability that the system is out of stock at time \( t \) is simply the probability that \( s + j \) or more units were demanded in the leadtime \( \tau \), since everything on order at time \( t - \tau \) arrives in the system by time \( t \) and nothing not on order at time \( t - \tau \) can arrive by time \( t \). Thus, averaging over \( j \),

\[
P_{\text{out}} = \frac{1}{Q} \sum_{j=1}^{\infty} P(s + j, \lambda \tau)
\]

where

\[
P(x, \lambda \tau) = \sum_{j=\infty}^{x} p(j, \lambda \tau); \quad p(j, \lambda \tau) = \frac{(\lambda \tau)^j}{j!} e^{-\lambda \tau},
\]

and \( p(j, \lambda \tau) \) is the Poisson probability that \( j \) units are demanded when the mean is \( \lambda \tau \). According to the Poisson properties \(^{(17)}\) it follows that

\[
P_{\text{out}} = \frac{1}{Q} \left[ \alpha(r) - \alpha(r+Q) \right]
\]

where

\[
\alpha(v) = \lambda \tau P(v, \lambda \tau) - v P(v+1, \lambda \tau).
\]
Finally, by the same sort of reasoning

\[(48)\quad B(Q,s) = \frac{1}{Q} \sum_{j=1}^{\infty} \sum_{y=j+1}^{\infty} (y-j-r) p(y,\lambda t),\]

and again making use of the Poisson properties, (17)

\[(49)\quad B(Q,s) = \frac{1}{Q} [\beta(s) - \beta(s+Q)]\]

where

\[(50)\quad \beta(v) = \frac{(\lambda t)^2}{2} P(v-1, \lambda t) - (\lambda t) v P(v, \lambda t) + \frac{v(v+1)}{2} P(v+1, \lambda t) .\]

The average annual cost is therefore

\[(51)\quad R = \frac{\lambda}{Q} A + IC \left[ \frac{Q}{2} + \frac{1}{2} + s - \mu \right] + \frac{\Pi}{Q} \lambda [\alpha(s) - \alpha(s+Q)] + \frac{\Pi+IC}{Q} [\beta(s) - \beta(s+Q)].\]

It is rather difficult to determine the optimal $Q,s$ values if it is necessary to use the exact cost expression, Eq. 51. The difficulty is caused by the $\alpha(s+Q), \beta(s+Q)$ terms. These terms will be important only if there is a considerable probability that when an order arrives, it will be insufficient to meet all the outstanding backorders. Such a situation could arise only if it did not cost much to incur backorders. In practice, it is generally quite costly to incur backorders and hence in practice it is usually a good approximation to ignore the $\alpha(s+Q)$ and $\beta(s+Q)$ terms. When this is true, it becomes a relatively simple task to determine numerically $Q^*,s^*$. If $Q$ is treated as continuous and $s$ as discrete, the necessary conditions which $Q^*,s^*$ must satisfy are

\[(52a)\quad Q = \left\{ \frac{2\lambda}{IC} \left[ A + \Pi \alpha(s) + \frac{\Pi+IC}{\lambda} \beta(s) \right] \right\}^{1/2}\]
\[(52b) \quad \left[1 - \frac{\lambda + IC}{\lambda \pi} \right] P(s, \lambda \tau) + \frac{\lambda + IC}{\lambda \pi} p(s, \lambda \tau) > \frac{QIC}{\Pi A} \]

A numerical procedure to determine \(Q^*, s^*\) is to set \(Q = Q\) in Eq. 52b and find the largest \(s\) which satisfies Eq. 52b -- calling it \(s_1\). Use \(s_1\) in Eq. 52a and solve for \(Q\) -- calling it \(Q_2\). Use \(Q_2\) in Eq. 52b and determine the largest \(s\) satisfying Eq. 52b -- calling it \(s_2\). Use \(s_2\) in Eq. 52a to determine \(Q_3\), etc. This iterative procedure usually converges rather rapidly.

An exact treatment of stochastic leadtimes in \((Q, s)\) models is very difficult because it is impossible to assume that leadtimes for different orders are independent random variables. If they were independent, then orders could cross. This is something that rarely happens in practice. If there is only a slight chance that two or more orders will interact, then one can carry out the above sort of analysis treating the leadtimes as independent random variables. This simply leads to the use of the marginal distribution of leadtime demand. For example, if the leadtime distribution can be approximated by a gamma distribution, then the marginal distribution of leadtime demand will be a negative binomial one. Consequently \(P_{out}\) and \(B(Q, s)\) will involve various combinations of negative binomial rather than Poisson terms. An exact solution to the lot-size, reorder point model has been obtained for Poisson demands and exponential leadtimes, when it is assumed that the leadtimes are independent.\(^{(18)}\) This model is not realistic from a practical point of view since an exponential distribution seldom represents leadtimes accurately, and, furthermore, orders cannot normally cross.

Exact equations for a lot-size, reorder-point model have not been worked out for the lost sales case when more than a single order can be
outstanding. The difficulty in working out this case arises from the need to take explicit account of the number of outstanding orders and the times that the orders were placed.

VARIATIONS OF THE \((q,s)\) MODELS

The \((q,s)\) models discussed in the previous section assumed that the unit cost of the item was constant, independent of the quantity ordered. Nonlinear variations of the unit cost with \(q\), such as quantity discounts, may also be considered. Usually, however, the computational effort required to determine the optimal \(q\) and \(s\) is increased considerably.

A frequently encountered type of quantity discount is that where the unit cost is \(C_0\) if \(0 < q < q_1\), \(C_1\) if \(q_1 \leq q < q_2\), \(C_2\) if \(q_2 \leq q \leq q_3\), etc., ..., \(C_m\) if \(q \geq q_m\). We can now imagine that \(m+1\) cost expressions are formulated \(R_0\), ..., \(R_m\), one for each value of the unit cost. If these are optimized over \(s\) for any \(Q\) so that each depends on \(Q\) alone, \(m+1\) curves are obtained, as shown in Fig. 6. Only the solid

![Diagram](image)

Fig. 6—Cost curve when quantity discounts are offered
portion of each curve is physically realizable. Note that these curves cannot cross. This discussion should show that the following computation scheme will yield the optimal \( q \) and \( s \) values.

First compute \( q_m^*, s_m^* \) for \( R_m \); if \( q^* \) satisfies \( q_m^* \leq q_m \), then \( q_m^* \) is optimal. If \( q_m^* < q_m \), compute \( R_m(q_m) \) and go on to the next stage. Now compute \( q_{m-1}^*, s_{m-1}^* \) for \( R_{m-1} \). If \( q_{m-1}^* \leq q_{m-1}^* < q_m \), compute \( R(q_{m-1}^*) \). If \( R(q_{m-1}^*) < R_m(q_m) \), then \( q_{m-1}^* \) is optimal; if \( R(q_{m-1}^*) > R_m(q_m) \), then \( q_m \) is optimal. When \( q_{m-1}^* \) is not physically realizable, compute \( R_{m-1}(q_{m-1}) \) and \( \hat{R}_{m-1}(\hat{q}_{m-1}) = \min\{R_{m-1}(q_{m-1}), R_m(q_m)\} \), where \( \hat{q}_{m-1} = q_m \) or \( q_{m-1} \).

In this case compute \( q_{m-2}^*, s_{m-2}^* \). If \( q_{m-2}^* \leq q_{m-2}^* < q_{m-1}^* \), compute \( R_{m-2}(q_{m-2}^*) \) and compare with \( \hat{R}_{m-1}(\hat{q}_{m-1}) \). If \( R_{m-2}(q_{m-2}^*) < \hat{R}_{m-1}(\hat{q}_{m-1}) \), \( q_{m-2}^* \) is optimal and if \( R_{m-2}(q_{m-2}^*) > \hat{R}_{m-1}(\hat{q}_{m-1}) \), \( \hat{q}_{m-1} \) is optimal. If \( q_{m-2}^* \) is not physically realizable, compute \( \hat{R}_{m-2}(\hat{q}_{m-2}) = \min\{R_{m-2}(q_{m-2}), \hat{R}_{m-1}(\hat{q}_{m-1})\} \). Repeat the above process. This iterative procedure will lead to the optimal solution. Equally effective computational procedures can be devised for handling other kinds of quantity discounts.

Another aspect of \((q,s)\) models not considered in the previous section was the matter of constraints. If the system stocked more than a single item, we treated the items as if there were no interactions between them. In practice, the items can interact in many ways. They may be competing for warehouse floorspace, for the allowable in-
vestment, for the number of possible procurements, etc. In theory, constraints of this sort can be handled by using Lagrange multipliers. The simplest way to illustrate the procedure is to consider just a single item. Suppose that the maximum inventory investment allowed for any given item is D. This means that

\[ C(Q+s) \leq D. \]

(53)

First the problem is solved ignoring the constraint. If the \( Q^* \), \( s^* \) so obtained satisfy Eq. 53, they are optimal. If not, the constraint is active: i.e., holds as a strict equality. Then a Lagrange multiplier \( \lambda \) is introduced and the function

\[ F = \frac{\lambda}{Q} A + IC (\frac{Q}{2} + s) - ICu + \Pi_{out} + (\hat{h} + IC) B(Q,s) + \Psi[D - C(Q+s)]. \]

The \( Q \) and \( s \) values which optimize \( F \) are found as a function of \( \lambda \), and \( \lambda \) is chosen so that \( Q,s \) satisfy Eq. 53 as a strict equality. The \( Q \) and \( s \) so obtained are optimal.

PERIODIC REVIEW MODELS

We have concerned ourselves in the preceding discussion with either deterministic or stochastic lot-size, reorder-point models. Now we will examine "periodic review" models, in which ordering decisions are made only at discrete, equally-spaced time intervals. The times at which the state of the system is reviewed will be called "review times." At each of these times, the decision maker must determine whether to order or not, and what the size of orders should be. In the time between two successive reviews, the system is said to have gone through one period's operation.
We will begin, as with stochastic \((Q,s)\) models, by studying two simple approximate formulations, one for the backorder and one for the lost sales case. It will be assumed that the operating doctrine for these models is to place an order at each review time for a sufficient quantity to bring the appropriate inventory level up to a predetermined value \(S\). In the mathematical development it will be assumed that all variables can be treated as continuous.

Consider first the backorder case. We wish to determine the optimal value of \(S\), the level to which the inventory position is returned at each review time, and \(T\), the optimal time between reviews. Let \(A\) be the cost of review and of placing an order; \(IC\) the cost of holding a unit in stock for one year (\(C\) being the unit cost); \(\mu\) the cost of a backorder (no time-dependent backorder cost is included in this simplified model); and \(\lambda\) the mean rate of demand. The average annual cost of placing orders is \(A/T\), and the expected annual cost of carrying inventory \(IC(S - \lambda T/2 - \mu)\), where \(\mu\) is the expected leadtime demand. This expression assumes that the expected annual unit years of shortage are negligible in comparison with the expected unit years of storage incurred per year. In short, integrating the expected net inventory over time yields essentially the same result as integrating the expected onhand inventory. To compute the expected number of backorders incurred per year we will first compute the expected number incurred per period. Suppose that an order is placed at time \(t_1\); the next order will be placed at time \(t_1 + T\). Assuming that no backorders will still be on the books after the arrival of the order placed at \(t_1\), it follows that for a given leadtime \(\tau\), the expected number of backorders incurred from \(t_1 + \tau\) to \(t_1 + \tau + T\) is

\[
\int_{S-S}^{\infty} (x-S) f(x,\tau+T) \, dx
\]
where \( f(x,t) \) is the density function for the quantity demanded in a time
\( t \). If the leadtime is a random variable with density \( g(\tau) \) and if
\[
\hat{h}(x,T) = \int_0^T f(x, \tau + T) g(\tau) \, d\tau ,
\]
then the expected number of backorders per period is
\[
\eta(S,T) = \int_S^\infty (x-S) \hat{h}(x,T) \, dx
\]
and the expected annual cost of backorders is \( \eta / T \). Note that whereas
for \((Q,s)\) models the expected backorders incurred per year depend on the
leadtime demand, in the periodic review case this quantity depends on
the leadtime plus the time between reviews. The expected annual cost
\( R \) is then
\[
R = \frac{A}{T} + IC \left[ S - \frac{\lambda T}{2\mu} \right] + \frac{\mu}{T} \int_S^\infty (x-S) \hat{h}(x,T) \, dx .
\]
The optimal \( S \) must satisfy the equation
\[
\frac{\partial R}{\partial S} = IC - \frac{\mu}{T} \hat{h}(S,T) = 0
\]
where \( \hat{h}(x,T) \) is the complementary cumulative of \( \hat{h}(x,T) \). Thus \( S \) for a
given \( T \) is determined from
\[
H(S,T) = \frac{IC T}{T} .
\]
To determine the optimal \( T \), the most practical procedure is to tabulate
the cost \( R \) as a function of \( T \), \( R \) being computed using the optimal value
of \( S \) in each case), and to determine in this way the \( T \) which minimizes \( R \).

The corresponding model for the lost sales case is obtained by
slightly modifying the inventory term in the backorder case. The expected amount on hand after the arrival of the order placed at time $t_1$ is $\lambda T$ (the expected amount ordered) plus the safety stock. The expected amount on hand at time $t_1 + \tau + T$ is the safety stock

$$S - \mu - \lambda T + \int_0^\infty (x-S) \hat{h}(x,T)dx = S - \mu - \lambda T + \eta(S,T),$$

where $S$ is the value of the onhand plus on order inventory level up to which one orders at the beginning of each period. Hence the average inventory is

$$S - \mu - \frac{\lambda T}{2} + \eta(S,T).$$

The above computations require that only a very small number of lost sales are incurred in any period. The expected annual cost for the lost sales case is then

$$R = \frac{A}{T} + IC(S - \mu - \frac{\lambda T}{2}) + \left(\frac{\pi}{2} + IC\right)\eta(S,T),$$

and the equation corresponding to Eq. 57 is

$$\hat{h}(S,T) = \frac{ICT}{\pi + ICT}.$$

Note that both the models just developed hold for an arbitrary number of orders outstanding.

The two models above assumed that an order was placed at each review time. It may not always be desirable to do this if, for instance, ordering costs are relatively high and review costs are relatively low. Of course, if review costs are low enough, one will use transaction recording and a $(Q,s)$ model. In practice, periodic
review models are frequently encountered, because it is too costly to operate a transactions-recording system. When review costs are high, one will normally wish to place an order at each review time, and an order up to $S$ policy will be essentially optimal.

There are several alternatives to the order up to $S$ policy which do not necessarily require the placement of an order at each review. One might employ a $(Q,k)$ or modification of the $(Q,s)$ model. Here, a level $k$ would be set, and if, at the review time, the inventory level is less than or equal to $k$, a quantity $Q$ is ordered. This policy is not very attractive, since the demand in a given period may bring the level below $k-Q$, and ordering $Q$ would not bring the level back even to the critical value $k$. A modification of this rule is to order an integral multiple of $Q$: i.e., a quantity $nQ$, where $n = 1, 2, ..., n$, is the largest positive integer, such that the inventory position after ordering lies between $k$ and $k + Q$. This policy will be referred to as an $nQ$ doctrine. Another possibility is to bring the inventory up to a level $K$ if at a review time it is below $k$. This is known as a $Kk$ policy, and a model using it is referred to as a $(K,k)$ model.

Let us now study a model in which an integral multiple of $Q$ is ordered if the inventory position is below $k$. More precisely, if at a review time the inventory position $y$ is equal to or less than $k$, an order for $nQ$ units is placed where $n$ is the largest integer, such that $y + nQ \leq k + Q$: i.e., $n$ is chosen to be as large as possible without bringing the inventory position above $k + Q$.

We shall restrict our attention to the backorder case with Poisson demands, constant leadtimes, and a unit cost that is independent of the quantity ordered. As before, the cost of a backorder will be assumed
to have the form \( I + \hat{A}t \), IC is the annual cost of carrying a unit in inventory, and \( A \) is the cost of placing an order. In addition, it will be assumed that making a review has a cost \( D(T) \), which may depend on the time between reviews \( T \).

Consider now the determination of the expected annual cost expression. Since there are an average of \( 1/T \) reviews per year, the average annual cost of making reviews is \( D(T)/T \). Let \( P_{or} \) be the probability of placing an order at any review. Then the expected annual fixed costs of placing orders will be \( AP_{or}/T \). The expected annual cost of backorders will be \( \sum_{q=1}^{Q} P_{or}E_{1}(q,k) + \hat{B}_{T}(q,k) \), where \( E_{T} \) is the expected number of backorders incurred per year and \( B_{T} \) is the expected number of backorders outstanding at any given point in time. Now let us evaluate \( P_{or} \), \( E_{T} \), and \( B_{T} \).

We have proved in Ref. 19 that the probability that the system is in state \( k + j, j + 1, \ldots, Q \), immediately after a review is \( 1/Q \). If the inventory position is \( k + j \) after one review, the probability that an order will be placed at the next review is \( P(j, \lambda T) \). Hence the probability of placing an order at any review period is

\[
P_{or} = \frac{1}{Q} \sum_{j=1}^{Q} P(j, \lambda T).
\]

From the Poisson properties, this can be written

\[
P_{or} = \frac{\lambda T}{Q} \left[ 1 - P(Q, \lambda T) \right] + P(Q+1, \lambda T).
\]

Next, \( E_{T}(q,k) \) will be evaluated. To calculate the expected number of backorders incurred per period, we will first compute the expected number incurred between \( T \) and \( T + T \), when the inventory position is \( k + j \) at the review time \( 0 \). To do this it is only necessary to determine
the expected number of backorders incurred from \( t = 0 \) to \( t = \tau + T \) and subtract from this the expected number incurred from \( t = 0 \) to \( t = \tau \).

This gives

\[
(63) \quad \sum_{v=k+1}^{\infty} (v-k-1) \left\{ p[v,\lambda(T+\tau)] - p(v,\lambda\tau) \right\}.
\]

Averaging Eq. 63 over the possible inventory positions, \( k + 1, \ldots, k + Q \), gives the expected number of backorders incurred per period, and multiplying by the average number of reviews per year, the following expression for \( E_T(Q,k) \) is obtained

\[
E_T(Q,k) = \frac{1}{QT} \sum_{u=k+1}^{k+Q} \sum_{v=u}^{\infty} (v-u) \left\{ p[v,\lambda(T+\tau)] - p(v,\lambda\tau) \right\}
\]

\[
(64) \quad = \frac{1}{T} [\hat{g}(k) - \hat{g}(k+Q)]
\]

where

\[
(65) \quad \hat{g}(x) = \frac{1}{Q} \sum_{u=x+1}^{\infty} \sum_{v=u}^{\infty} (v-u) \left\{ p[v,\lambda(T+\tau)] - p(v,\lambda\tau) \right\}.
\]

By use of the Poisson properties (17) it can be demonstrated that

\[
(66) \quad \hat{g}(x) = \alpha[x,\lambda(T+\tau)] P[x,\lambda(T+\tau)] + \beta[\lambda(T+\tau)] P[x-1, \lambda(T+\tau)]
\]

\[
+ \gamma[\lambda(T+\tau)] P[x-2, \lambda(T+\tau)] - \alpha(x,\lambda\tau) P(x,\lambda\tau)
\]

\[
- \beta(\lambda\tau) P(x-1,\lambda\tau) - \gamma(\lambda\tau) P(x-2,\lambda\tau)
\]

where

\[
(67) \quad \alpha[x,\lambda\tau] = \frac{1}{Q} \left[ \lambda\tau(1-x) + \frac{x(x+1)}{2} \right],
\]

\[
(68) \quad \beta[\lambda\tau] = \frac{\lambda\tau(\lambda\tau-1)}{Q},
\]

\[
(69) \quad \gamma[\lambda\tau] = -\frac{(\lambda\tau)^2}{2Q}.
\]
The expected stockout term in this form is fairly easy to handle computationally.

It remains to calculate the \( B_T(Q,k) \) term. When the inventory position is \( k + j \) at time 0, the expected unit years of shortage from \( \tau \) to \( \tau + T \) are

\[
\int_\tau^{\tau+T} \sum_{u=k+j}^\infty (u-k-j) p(u,\lambda t) dt.
\]

Averaging over the \( Q \) possible inventory positions and converting to a yearly basis gives

\[
B_T(Q,k) = \frac{1}{Q T} \sum_{u=k+1}^{k+Q} \int_\tau^{\tau+T} \sum_{v=u}^\infty (v-u) p(v,\lambda t) dt.
\]

Using the Poisson properties, (17) one obtains

\[
B_T(Q,k) = J(k,\lambda T) - J(k+Q,\lambda T) - J[k,\lambda(T+\tau)] + J[k+Q,\lambda(T+\tau)]
\]

where

\[
J(x,\lambda t) = \frac{\lambda t x^2}{2Q} P(x,\lambda t) - \frac{\lambda t^3}{6Q} P(x-1,\lambda t) - \frac{x(x+1)t}{2Q} P(x+1,\lambda t)
\]

\[
+ \frac{x(x+1)(x+2)}{6Q^2T} P(x+2,\lambda t).
\]

The average inventory position of the system in a period in which the initial position is \( k + j \) is

\[
\frac{1}{T} \int_0^T \sum_{u=0}^\infty (k+j-u) p(u,\lambda t) dt = k + j - \frac{\lambda T}{2}.
\]

Averaging over the \( Q \) initial states, one obtains the following as the expected inventory position at any point in time.

\[
\frac{1}{Q} \sum_{j=1}^Q (k+j-\frac{\lambda T}{2}) = k + \frac{Q+1}{2} - \frac{\lambda T}{2}.
\]
The expected inventory on hand plus on order is obtained by adding average backorders to Eq. 75. Subtracting expected leadtime demand, μ, gives the onhand inventory.

Combining the terms derived above into an expression for expected annual variable costs gives the following:

\[
R_T(Q, k) = \frac{D(T)}{T} + \frac{A}{T} \bar{P}_{or} + IC\frac{Q+1}{2} + k - \frac{\lambda T}{2} - \mu + B_T(Q, k)
\]

\[
+ \bar{R}_k(Q, k) + \bar{B}_T(Q, k).
\]

By minimizing Eq. 76 it is possible to find optimal values of Q, k for a given T or to find optimal values of Q, k, and T. The task of making numerical computations with this model is quite difficult and it is usually necessary to make use of a digital computer to determine \(Q^*\) and \(k^*\).

It is interesting to note that the formulas for the transactions-recording \((Q, s)\) model can be obtained from those of the above periodic review model by taking the limit as the review interval tends to zero and dropping the \(\frac{D(T)}{T}\) term which becomes infinite.

A special case of the above model which is of particular interest is that where \(Q = 1\). When \(Q = 1\), an order will be placed at each review time provided that there have been any demands at all in the preceding period. Thus on setting \(Q = 1\), the exact equations for the simple model treated at the beginning of this section obtain in situations where the demand is Poisson distributed. If \(S = k+1\), then on setting \(Q = 1\) in Eq. 76, it follows that the average annual cost of the "order up to S" policy is
\[ R_T(S) = \frac{D(T)}{T} + A T P(1, \lambda T) + IC\left[S - \mu \frac{\lambda T}{2}\right] + \tau E(1, S-1) + (IC + \hat{\gamma})B(1, S-1) \]

The optimal value of $S$ is then the largest integer $S$ for which

\[ IC = \frac{1}{T} \left[ \tau - \frac{S}{\lambda} (IC + \hat{\gamma}) \right] \left\{ P(S, \lambda(\tau + T)) - P(S, \lambda T) \right\} \]

\[ - \left( \frac{IC + \hat{\gamma}}{T} \right) \left\{ (\tau + T)P(S, \lambda(\tau + T)) - \tau P(S, \lambda T) \right\} \]

\[ - \left( \frac{IC + \hat{\gamma}}{\lambda T} \right) \left\{ P(S, \lambda(\tau + T)) - P(S, \lambda T) \right\} < 0. \]

If $\tau = 0$, the optimal $S$ is the largest one for which

\[ P(S, \lambda T) > \frac{IC}{\hat{\gamma} + IC}. \]

One would expect that the optimal solution to Eq. 76 would yield $Q^* = 1$, when the review costs were high in comparison to the ordering costs. Surprisingly, $Q^* = 1$ under more conditions than one might anticipate. For practical purposes, it seems that $Q^*$ almost always $= 1$, and hence that the simpler model represented by Eq. 77 can be used in place of that whose cost expression is given by Eq. 76.

We must still consider the $(K, k)$ models. Note that a periodic review system is using a $Kk$ operating doctrine if an order is placed at a review time only when the inventory position (for the backorders cost) is less than or equal to $k$. If this is true, then sufficient quantity is ordered to bring the inventory up to level $K$. As before, we assume that units are requested singly and that the demand during any period is Poisson distributed. Let us define a "cycle" of the system as the time existing be-
tween the placement of two successive orders. If we compute the expected cost per cycle and then multiply by the average number of cycles per year, we will obviously have the average annual cost.

For a periodic review system, a cycle may consist of 1, 2, 3, \ldots periods. Suppose that an order is placed at time \( \hat{t} \). Imagine that at the review time \( \hat{t} + nT \), the inventory position of the system is \( k + j \) and no order has been placed since \( \hat{t} \). The probability of this is \( p(K-k-j, n\lambda T) \). Denote by \( H(k+j, T) \) the expected cost of carrying inventory and of backorders incurred from \( \hat{t} + nT + \tau \) to \( \hat{t} + (n+1)T + \tau \).

The expected cost per cycle of carrying inventory and of backorders is then

\[
\sum_{n=0}^{\infty} \sum_{j=1}^{K-k} p(K-k-j, n\lambda T) H(k+j, T)
\]

where by definition \( p(K-k-j, 0) = 0, K-k-j \neq 0, p(0, 0) = 1 \).

Next, we will determine the expected length of a cycle. This is simply \( T \) times the expected number of periods contained in one cycle. A cycle will contain precisely one period if the demand is greater than or equal to \( K-k \) in one period. The probability of this is \( P(K-k, \lambda T) \).

The probability that precisely \( n (n \geq 2) \) periods are included in a cycle is

\[
\sum_{j=1}^{K-k} \sum_{n=1}^{K-k} p[K-k-j, (n-1)\lambda T] P(j, \lambda T)
\]

Thus the expected length of a cycle is

\[
T \sum_{n=1}^{K-k} \sum_{j=1}^{K-k} n p[K-k-j, (n-1)\lambda T] P(j, \lambda T)
\]

If \( D \) is the review cost and \( A \) the cost of placing an order, the average annual cost is
\[
R(K,k,T) = \frac{D}{T} + \frac{A + \sum_{n=0}^{\infty} \sum_{j=1}^{\infty} p(K-k-j,n\lambda T) H(k+j,T)}{\sum_{n=1}^{K-k} \sum_{j=1}^{\infty} n p[K-k-j,(n-1)\lambda T] P(j,\lambda T)}
\]

The explicit form for \(H(k+j,T)\) may be easily obtained from what was derived above for the \(nQ\) model. There seems to be no easy way to convert the terms in Eq. 82 into a form which does not involve a sum over \(n\). It would be out of the question to attempt to determine manually the optimal values of \(K\), \(k\), and \(T\). However, the job can be carried out on a digital computer without too much difficulty.

**ALTERNATIVE STOCKOUT PROCEDURES**

The models discussed in the previous sections assumed that incurring backorders or lost sales resulted in a cost of some sort. Such costs are often rather hard to determine in the real world. An alternative procedure often used in steady-state models is to specify the probability of being out of stock at any point in time, i.e., \(P_{out}\). This substitute method is really included in the formulations of the models which assign costs to stockouts, as will now be shown.

To be specific, imagine that the system is to be operated using a \(Qs\) doctrine. Instead of specifying a stockout cost, imagine that it is required that the probability of being out of stock at any time is to be no greater than \(\alpha\). It is desired, then, to determine \(Q\) and \(s\) such that the expected annual cost of ordering, and holding inventory is minimized subject to the constraint. It is clear that the constraint will hold as a strict equality. Hence it is desired to minimize

\[
\frac{\lambda}{Q} A + IC[\frac{Q}{2} + s - \mu + B(Q,s)]
\]
subject to the constraint

\[ P_{\text{out}} = \Omega. \]

From the theory of Lagrange multipliers it is clear that if we introduce a Lagrange multiplier, which will be written \( \Pi \lambda \), the above constrained minimization problem is equivalent to the unconstrained problem where it is desired to minimize

\[ R(Q, s) = \frac{\lambda}{Q} A + IC \left[ \frac{Q}{2} + s - \mu + B(Q, s) \right] + \Pi \lambda P_{\text{out}} \]

where \( \Pi \) is selected so that \( P_{\text{out}} = \Omega \). This is precisely Eq. 51 with \( \hat{\Pi} = 0 \). Therefore specifying the probability of being out is equivalent to setting a value of \( \Pi \) and having \( \hat{\Pi} = 0 \).

It is equally clear that if instead a constraint is placed on the expected number of backorders at any point in time, this is equivalent to setting \( \Pi = 0 \) in Eq. 51 and thereby determining a value of \( \hat{\Pi} \).

**USE OF FUNCTIONAL EQUATIONS TO STUDY NATURE OF OPTIMAL OPERATING DOCTRINES**

In preceding sections, we developed a number of models which minimized average annual costs for a specified type of operating doctrine. We made no attempt to prove that any of these operating doctrines was the optimal one to use. We simply selected what seemed intuitively to be the preferred or most reasonable ones. To determine the structure of operating doctrines for a periodic review and transactions reporting systems, we must use a different approach: i.e., the functional equation technique of dynamic programming. The first paper in English to use this method was by Arrow, Harris, and Marschak. (14)
To begin, we will illustrate the procedure as it is used for periodic review systems. Consider the system at the time of a review \( t \). Let \( T \) be the time between reviews; \( \tau \) the procurement leadtime (here assumed to be constant); \( A \) the fixed cost of placing an order; and \( D \) the review cost. Let us imagine also that demands that occur when the system is out of stock are backordered. The inventory position of the system at time \( t \) (prior to the placing of an order) will be denoted by \( \xi \). The quantity ordered, if any, will be denoted by \( Q \geq 0 \). Let \( r(\xi + Q) \) be the expected cost of carrying inventory and of backorders incurred in the period \( t + \tau \) to \( t + \tau + T \) discounted to time \( t \): in short, the expected costs for a period displaced by a leadtime and discounted to time \( t \). Finally, at any review time \( t \): let \( \Lambda(\xi) \) represent the discounted expected cost over all future time when the inventory position at time \( t \) prior to the placing of any order is \( \xi \), and when an optimal ordering policy is followed at time \( t \) and at all future review times. Note that for a fixed \( \xi \), \( \Lambda(\xi) \) will be independent of the particular review time under consideration. In computing \( \Lambda(\xi) \), the convention will be used that the carrying and backorder costs which are associated with any review time are those incurred in a period displaced by a leadtime from the review time.

It is possible to relate \( \Lambda \) at any given review time \( t \) to \( \Lambda \) at \( t + T \). If a quantity \( Q \) is ordered at time \( t \), the costs associated with the period \( t \) to \( t + T \) will be

\[
D + A \delta + CQ + r(\xi + Q)
\]

where
(85) \[ \delta = \begin{cases} 0 & \text{if } Q = 0 \\ 1 & \text{if } Q > 0 \end{cases} \]

and \( \xi \) is the inventory position at time \( t \) before the placing of any order, provided that the unit cost of the item is a constant. If \( x \) units are demanded between \( t \) and \( t + T \), the inventory position at the next review time before the placement of any order will be \( \xi + Q - x \). If an optimal policy is followed at time \( t + T \) and all future times, the present worth of all future costs discounted to \( t + T \) will be \( \Lambda(\xi + Q - x) \). When we treat the demand as continuous, with density \( f(x) \), the probability that the demand between \( t \) and \( t + T \) lies between \( x \) and \( x + dx \) is represented by \( f(x)dx \). It must, therefore, be true, that \( \Lambda(\xi) \) satisfies the functional equation

\[
(86) \quad \Lambda(\xi) = \min_Q \left[ D + A\delta + CQ + r(\xi+Q) + \alpha \int_0^\infty f(x) \Lambda(\xi + Q - x)dx \right],
\]

where \( \alpha \) discounts to \( t \) costs known at time \( t + T \). An analysis of this functional equation makes it possible to obtain information concerning the nature of the optimal operating doctrine for the periodic review system under consideration.

Let

\[
(87) \quad F(y) = Cy + r(y) + \alpha \int_0^\infty f(x) \Lambda(\xi - x)dx.
\]

Then Eq. 86 can be written

\[
(88) \quad \Lambda(\xi) = \min_Q \left[ D + A\delta - C\xi + F(\xi+Q) \right]
\]

\[
= \min \left\{ D - C\xi + F(\xi) \quad \begin{cases} D + A - C\xi + \min_Q F(\xi+Q) \end{cases} \right. \]
Equation 88 reveals that the nature of the optimal operating doctrine depends only upon $F(y)$. Let us suppose that $F(y)$ has the shape shown in Fig. 7.

![Diagram](image)

Fig. 7 — Case where $F(y)$ has a single relative minimum

Then

$$\min F(\xi + Q) = \begin{cases} F(K), & \xi \leq k \\ Q, & \xi > k. \end{cases}$$

Therefore, the optimal policy has the form

$$Q = \begin{cases} 0, & \xi \geq k \\ K - \xi, & \xi < k. \end{cases}$$

In other words, an optimal policy is a Kk policy. When the system is operated optimally, the inventory position will never go above K.
If we suppose, on the other hand, that $F(y)$ has the shape shown in Fig. 8, then the optimal policy has this form:

![Graph of $F(y)$ with minima at $k_1$, $k_2$, and $k_3$.]

**Fig. 8—A case where $F(y)$ has two relative minima**

Order $K - \xi$, if $\xi < k_1$; do not order if $k_1 \leq \xi \leq k_2$; order $K - \xi$ if $k_2 < \xi < k_3$; do not order if $\xi \geq k_3$. Here, a $K_k$ policy is not optimal.

It is interesting to inquire about the circumstances under which a $K_k$ policy is optimal. This question has not yet been answered completely even though a large number of papers (1), (14 - 16), (20) have been devoted to the subject. Most papers make such restrictive assumptions (i.e., assuming $A = 0$ and a zero leadtime) that they are of little interest from a practical point of view. However, the recent
paper by Scarf (20) obtained results that were sufficiently general to be of practical interest. Scarf proved that if \( r(y) \) is convex, then a \( Kk \) policy will be optimal. The types of costs that we have considered previously usually involve a convex \( r(y) \) (when all variables are treated as continuous); hence a \( Kk \) policy will normally be the optimal one for periodic review systems. Ordering up to \( S \), or ordering in multiples of a fundamental unit \( Q \), are only approximations to the optimal policy. For practical purposes, however, the "order up to \( S \)" procedure (really a special \( Kk \) policy, with \( S = K \) and \( S = k + 1 \) for discrete variables) seems optimal most of the time -- even when review costs are relatively small compared with the ordering cost. Really, it is surprising that this simple doctrine is essentially optimal under such a wide range of conditions. It should be emphasized again, however, that the discussions above all assumed that the unit cost was constant. When quantity discounts are available, it is no longer necessarily true that a \( Kk \) policy is optimal.

Transactions-recording systems can be submitted to the same sort of analysis as that used above. A convenient way to begin is to consider the system immediately after a demand has occurred but before any decision has been made as to whether or not to place an order.

Let \( \Lambda (\xi) \) be the present value of all future costs at the time a demand occurs, if \( \xi \) is the inventory position before the placing of any order and if an optimal policy is followed after the occurrence of every demand. In computing, the carrying and backorder costs will be displaced by a leadtime. Let us assume that a demand occurs at time 0. If \( g(t) \) is the density function for the time between demands, the probability that the next demand will occur between \( t \) and \( t + dt \) is
\text{g(t)dt. It is being assumed here that g depends only on t and not}
on any past history of the process. We are concerned here with a
situation in which units are demanded one at a time, but the analysis
can be generalized to apply where the order size is a random variable,
as in (13) or (21). Continuous discounting will be used, so that the
discount factor has the form $e^{-it}$. We are assuming that a demand oc-
curs at time 0; that the inventory position of the system is $\xi$ before
any order is placed; and that $Q \geq 0$ is the quantity ordered. Let
\( r(\xi + Q) \) be the expected cost discounted to time 0 of carrying inven-
tory and of backorders from time $\tau$ (the leadtime) to $\tau + t$ (t being
the time when the next demand occurs). Then it is clear that $\Lambda(\xi)$
must satisfy the following functional equation:

\begin{align*}
(90) \quad \Lambda(\xi) &= \min_Q \left[ A\xi + CQ + r(\xi + Q) + \left\{ \int_0^\infty e^{-it} g(t)dt \right\} \Lambda(\xi + Q - 1) \right] \\
\end{align*}

if the unit cost of the item is constant. Now, let

\begin{align*}
(91) \quad \alpha &= \int_0^\infty e^{-it} g(t)dt. \\
\end{align*}

Then Eq. 90 can be written

\begin{align*}
(92) \quad \Lambda(\xi) &= \min_Q \left[ A\xi + CQ + r(\xi + Q) + \alpha \Lambda(\xi + Q - 1) \right] \\
\end{align*}

This form is very close to that of Eq. 86. If $r(\xi + Q)$ is convex, a
Kk policy will be optimal. Since we have assumed here that units are
demanded individually, it follows that a Kk policy is equivalent to a
Q\text{=} policy. One can immediately obtain from this the fact that a Qs
policy is indeed the optimal one for the simple models studied under
"Deterministic Models with a Constant Rate of Demand."
Let us assume now that one is studying a periodic review or transactions recording system for which the optimal policy is known to be a \(Kk\) policy. The functional equations can be used to obtain the average annual cost expression for either of these systems. First, however, the Eqs. 86 and 92 should be reformulated so that the variable refers to the inventory position of the system after any order has been placed. Suppose for the periodic review system that \(Y(\xi; K, k)\) represents the discounted cost at a review time after any order has been placed, \(\textbf{if}\) a \(Kk\) operating doctrine is used (\(K\) and \(k\) being the values -- not necessarily the optimal ones -- of the critical levels). Then \(Y(\xi; K, k)\) must satisfy the functional equation

\[
y(\xi; K, k) = r(\xi) + \alpha \int_{0}^{\xi-k} f(x) Y(\xi-x; K, k) \, dx
\]

\[
+ \alpha \int_{\xi-k}^{\infty} f(x) [A + C(\xi-x) + Y(\xi; K, k)] \, dx.
\]

The corresponding equation for the transactions recording system when units are demanded one at a time and when a \(Qs\) operating doctrine is used becomes

\[
Y(\xi; Q, s) = \begin{cases} 
\tau(\xi) + \alpha Y(\xi-1; Q, s), & \xi > s + 2 \\
\alpha(A + CQ) + r(s+1) + \alpha Y(Q+s; Q, s), & \xi = s+1.
\end{cases}
\]

Written out in more detail Eq. 94 becomes

\[
Y(Q+s; Q, s) = \tau(Q+s) + \alpha Y(Q+s-1; Q, s) \\
Y(Q+s-1; Q, s) = \tau(Q+s-1) + \alpha Y(Q+s-2; Q, s) \\
\vdots
\]

\[
Y(s+1; Q, s) = bA + \alpha CQ + \tau(s+1) + \alpha Y(Q+s; Q, s).
\]
Thus
\[
Y(Q+s; Q, s) = \sum_{j=0}^{Q-1} \alpha^j r(Q+s-j) + \sigma Q A + \sigma Q C Q + \sigma Q Y(Q+s; Q, s)
\]

or
\[
Y(Q+s; Q, s) = (1 - \alpha Q)^{-1} \sum_{j=0}^{Q-1} \alpha^j r(Q+s-j) + (1 - \alpha Q)^{-1} [\sigma Q(A + C Q)],
\]

and an explicit expression for \(Y(Q+s; Q, s)\) has been obtained. In the periodic review case, too, it is also possible to solve explicitly for \(Y(K; K, k)\) when the demand is treated as discrete, but the task is more difficult (see Ref. 13). Now, the interesting thing is that one can demonstrate (13) that the average annual cost \(R\) for the periodic review system is given by
\[
R = \lim_{t \to 0} Y(K; K, k)
\]

and the average annual cost \(R\) for the transactions recording system is given by
\[
R = \lim_{t \to 0} Y(Q+s; Q, s)
\]

where \(t\) is the interest rate which appears in the discounting factor \(\sigma\) and in \(r(t)\). The functional equation approach therefore provides another way of obtaining the average annual cost expressions. As we have noted earlier, it can also be used to determine the nature of the optimal operating doctrine.

It is of no assistance, however, in studying the lost sales case. In the latter, one cannot generally have the functional recurrence relation involve only a single parameter. Instead, \(Y\) must also be a
function of the quantities ordered at each previous review over the span of a leadtime. This seems to eliminate the possibility of characterizing in any simple way the resulting functional equation.

This will complete the discussion of steady state models in which the mean rate of demand remains constant. The next section will summarize the mathematical techniques that have been used to study such models.

SUMMARY OF MATHEMATICAL TECHNIQUES USED IN THE STUDY OF STEADY STATE STOCHASTIC MODELS

A variety of mathematical techniques have been employed in developing the models studied above. All of the models themselves, however, had one common feature: in one way or another, each could be represented as a Markov process that was either discrete or continuous in time. The characteristic property of Markov processes is their making it possible to define a set of states for the system such that at every point in time or at certain discrete points in it, when observation takes place, the system will be in one of the states. Furthermore, if the system is in a given state \( i \) at time \( t \), the probability that it will be in some future state \( j \) at time \( t + \tau \) depends upon \( i \) and \( j \) but not upon the history of the system: i.e., how it came to be in state \( i \) at time \( t \). Essentially nothing has been done in treating inventory systems which cannot be represented as a Markov process.

Inventory systems are characterized by a process of renewal. Typically, at certain points in time events occur which restore inventory systems to previous states (i.e., the placement of an order restores the inventory position to the level \( K \), for instance). A system is said
to go through a cycle between two successive renewal events. The reader will recall that in previous sections a recurring technique for determining the average annual cost was to compute the expected cost per cycle and then to multiply by the average number of cycles per year. This is a general technique which can be used both for periodic review and for transactions recording systems. It can in fact be used to obtain the average annual cost for transactions reporting systems when the order size is a random variable. We have not illustrated this process here simply because the equations become quite complicated.

Toward the end of our discussion, "Lot-Size, Reorder-Point Models with Stochastic Demands," we introduced another approach to obtaining the average annual cost. This method was to determine directly the state probabilities (here, using queuing theory) and then to obtain the average annual cost by averaging over the state probabilities. The queuing technique becomes especially useful when the leadtimes are treated as independent random variables. A procedure that was used with periodic review models was to determine the expected cost per period and then multiply by the average number of periods per year. The final approach introduced was to obtain the average annual cost from the present worth of all future costs (i.e., using the functional equation method).

The lost sales case represents a fundamentally different stochastic process from the backorder case, which seems to be much more difficult to treat analytically. Very little analysis of this case has been carried out. It is much easier to obtain useful results for the backorders case.
In previous sections, we made no attempt to allow the time-dependent part of the backorder cost to be a nonlinear function of time. For transactions recording systems in which units are demanded singly, it is possible to obtain the average annual cost when the backorder cost is a nonlinear function of time. It is somewhat more difficult to do this for periodic review systems, but this has also recently been worked out by one of the authors.

When the order size can be a random variable, the backorder cost could also be a nonlinear function of the number of backorders. Nothing seems to have been done in studying such nonlinearities, perhaps because so little attention has been directed at systems where the order size can be a random variable. The introduction of nonlinear stockout penalties into inventory models is not currently of great practical interest, since insufficient knowledge of the nature of backorder costs makes it impossible to specify the nonlinear cost function.

Nonlinear costs of the units in the form of price breaks have been considered in the previous sections. We made no attempt to consider cases where the fixed ordering cost is a nonlinear function of the number of orders placed. This sort of behavior can usually be accounted for by assuming a value of $A$. If the $N$ (average number of orders placed per year) obtained as a result of solving the model does not correspond to the appropriate $N$ for that $A$, a new $A$ is selected, and the problem is solved again.

Nor have we made an attempt to include nonlinear holding costs. Theoretical work along these lines is negligible, and very little information is available from a practical point of view to indicate whether or not such generalizations would be of value.
STATIC ONE-PERIOD MODELS

Some of the earliest inventory models involving probabilistic demand behavior were simple, one-period models such as the "newsboy" problem. Faced with a stochastic demand for papers that cost the newsboy m cents and sell for n cents per copy, the newsboy must decide on an initial stock $S$ of papers that maximizes his expected daily profit, $G(S)$. The expression for expected profits may be written

$$G(S) = -Sm + n \sum_{x=0}^{S} x p(x,\lambda) + nS \sum_{x=S+1}^{\infty} p(x,\lambda),$$

where $p(x,\lambda)$ is the probability that $x$ papers are demanded when the average daily demand is $\lambda$. The three terms, from left to right, represent the cost of the $S$ papers, the expected revenues received when $S$ or less papers are demanded, and the expected revenues when more than $S$ demands occur. The expected profits attributable to the $S^{th}$ paper may be obtained by subtracting $G(S-1)$ from $G(S)$. The condition for positive expected profits resulting from the $S^{th}$ paper may be written

$$G(S) - G(S-1) = -m + n \sum_{x=S}^{\infty} p(x,\lambda) > 0,$$

or

$$P(S,\lambda) > \frac{m}{n}, \quad P(S,\lambda) = \sum_{x=S}^{\infty} p(x,\lambda).$$

The largest positive integer $S$ that satisfies Eq. 101 will result in maximum expected profits to the newsboy. In the event that the random demand can be approximated with a continuous probability distribution, Eq. 101 can be replaced by
(102) \[ \int_{S}^{\infty} f(x, \lambda) dx = \frac{m}{n} \]

where \( f(x, \lambda) \) is the probability density function for daily demand.

The model is valid for any arbitrary distribution of demand.

Another type of single period inventory model of particular interest to the military is referred to as a flyaway kit problem. (13)(22)

Imagine that during a period of time of given length, demands can be incurred for \( n \) different items. A warehouse of fixed size is available. Any quantity desired of each of the items may be purchased provided that the total quantity acquired does not overflow the warehouse. No additional items can be purchased during the period over which demands are incurred. The only relevant cost is of being out of stock when a demand occurs. Let \( \Pi_{i} \) be the cost per demand for item \( i \) when the system has no stock of \( i \) on hand. The problem is to determine the quantity \( x_{i} \) to procure and place in the warehouse for each of the \( n \) items so as to minimize the expected shortage costs while not exceeding the capacity of the warehouse.

Let \( f_{i}(y_{i}) \) be the probability density for the number of units of \( i \) that will be demanded in the period. Assume that \( v_{i} \) is the unit volume of \( i \) and \( V \) is the volume of the warehouse. First imagine that the \( x_{i} \) can be treated as continuous. Then the problem is to minimize the expected shortage costs

(103) \[ \Gamma = \sum_{i=1}^{n} \Pi_{i} \int_{x_{i}}^{\infty} (y_{i} - x_{i}) f_{i}(y_{i}) dy_{i} \]
subject to the constraint

\[ \sum_{i=1}^{n} v_i x_i = V. \]

Introducing a Lagrange multiplier \( \gamma \), one sees that necessary conditions that the \( x_i \) must satisfy are

\[ \frac{\partial \gamma}{\partial x_i} + \gamma v_i = 0, \quad i=1, \ldots, n, \]

or

\[ \frac{\Pi_i}{v_i} F_i(x_i) = \frac{\Pi_j}{v_j} F(x_j) = \gamma, \quad \text{all } i, j, i \neq j \]

where \( F_i(x_i) \) is the complementary cumulative function for \( f_i(x_i) \). A numerical procedure is to select a value of \( \gamma \) arbitrarily and compute a set of \( x_j \) corresponding to this \( \gamma \). If the constraint is not satisfied (to the desired degree of accuracy), select a new \( \gamma \) and repeat the procedure.

The technique just described cannot handle integrality of demand in a satisfactory way, and cannot be used if only one or two units of each item are stocked. An alternative procedure which can exactly account for the integrality is dynamic programming. Now the probability that precisely \( x_i \) units will be demanded will be written \( p_i(x_i) \).

Let

\[ y_j(\xi) = \min_{x_j} \left[ \sum_{i=1}^{J} \Pi_i \sum_{y_i=x_i}^{\infty} (y_i-x_i)p_i(y_i) \right] \]

subject to

\[ \sum_{i=1}^{J} v_i x_i \leq \xi. \]
Then the recurrence relations used to compute the functions $\gamma_j(\xi)$ sequentially are:

\[
(105) \quad \gamma_j(\xi) = \min \left[ \prod_j \sum_{x_j}^{\infty} (y_j-x_j)p_j(y_j) + \gamma_{j-1}(\xi-x_j) \right], \quad j = 1, \ldots, n,
\]

where $\gamma_0(\xi) = 0$ for all $\xi$. The minimum cost is $\gamma_n(V)$ and the optimal quantities to procure are those corresponding to $\gamma_n(V)$.

Normally, a computer would be used to carry out the above computations.

It should be noted that dynamic programming need not give absolutely the best answer, since the shape of the items is as important in packing them in the warehouse as is their volume. A simple approximate procedure for solving the problem when integrality of demand must be accounted for is to select units one at a time in such a way that the incremental decrease in stockout cost per unit volume of floor-space required is maximized.

Typical practical applications of the flyaway kit model lie in stocking a submarine with spare parts for a cruise or loading an airplane for periodic replenishment of an isolated air base.

**DYNAMIC MODELS**

It is usually true in practice that the stochastic processes associated with an inventory system will change with time. When the change is sufficiently rapid to make unsuitable a steady state approximation, the task of studying the system analytically becomes much more difficult. With dynamic systems, it is not possible to obtain explicit analytical solutions for the quantities of interest. About
the best that can be done is to obtain numerical solutions for special cases. Sometimes, however, useful information can be obtained from a theoretical analysis, even if it is impossible to arrive at explicit analytical solutions. Various numerical techniques have been used to study dynamic inventory problems. The most frequent procedure, however, is to apply the recursive computational procedure of dynamic programming.

Unlike steady state models, dynamic ones do not adapt themselves to the use of an infinite planning horizon in making the computations. The numerical effort required to solve the problem increases, of course, as the length of the planning horizon increases. Fortunately, the distant future has little influence upon current decisions, and it often turns out that a fairly short planning horizon is satisfactory. It frequently happens that dynamic inventory situations are really combinations of inventory and production scheduling problems. This will be reflected in the models to be described in the two sections which follow. Initially, two deterministic dynamic models will be examined.

**DETERMINISTIC DYNAMIC MODELS**

Deterministic dynamic models often serve as useful tools for planning purposes. An extremely simple model of the combined inventory-production scheduling type might be described as follows. A company produces a certain item which is subject to a fluctuating demand pattern. The item is produced continuously. At certain times, however, the company may find it either necessary or desirable to employ overtime production. In addition, raw materials costs may fluctuate. The item can be stored as long as necessary. Because of the orders
Currently, it has been found possible to make some predictions concerning future demand for the item. These predictions become progressively less accurate for the distant future. The costs are known for producing a unit on either regular or overtime and for storing a unit for any given period (a day, a week, a month, etc.).

A planning horizon of $N$ periods is used. The company wishes to determine the production schedule which will meet exactly the predicted demands over the planning horizon and maintain in inventory a specified quantity at the end of the planning horizon, at the same time minimizing the sum of production and carrying costs.

To formulate the problem mathematically, let $x_{i,j}$ be the number of units produced on regular time in period $i$ to meet the demand in period $j$, and let $y_{i,j}$ be the number of units produced on overtime in period $i$ to meet the demand in period $j$. The regular and overtime capacities in period $i$ will be denoted by $R_i$ and $E_i$ respectively, and the predicted demand in period $i$ by $D_i$. Similarly, let $R_0$ be the initial inventory and $x_{0,j}$ the quantity of the initial inventory used to meet the demand in period $j$. Also, let $D_{N+1}$ be the desired onhand inventory at the end of period $N$, and $x_{i,N+1}, y_{i,N+1}$ the production on regular and overtime, respectively, in period $i$ for ending inventory, with $x_{0,N+1}$ the quantity initially in inventory which is still in inventory at the end of the planning horizon.

Finally, let $c_{i,j}$ be the cost of producing one unit on regular time in period $i$ and storing it for use in period $j$; let $d_{i,j}$ be the cost of producing one unit in period $i$ on overtime and storing it for use in period $j$; and let $c_{0,j}$ be the cost of storing one unit initially in inventory for use in period $j$. 

The constraints on the problem then take the form

\[(106) \quad \sum_{j=1}^{N+1} x_{ij} \leq R_i, \quad i = 0, 1, \ldots, N\]

and

\[(107) \quad \sum_{j=1}^{N+1} y_{ij} \leq E_i, \quad i = 1, \ldots, N,\]

which are the capacity constraints, and

\[(108) \quad \sum_{i=0}^{J} x_{ij} + \sum_{i=1}^{N} y_{ij} = D_j, \quad j = 1, \ldots, N + 1\]

which are the constraints that require the predicted demand be met exactly. Note that anything sold in period \( j \) must have been produced in period \( j \) or in an earlier period. It must also be true, of course, that the \( x_{ij} \) and \( y_{ij} \) be non-negative. The objective function which should be minimized is

\[(109) \quad \sum_{i=0}^{N} \sum_{j=1}^{N+1} c_{ij} x_{ij} + \sum_{i=1}^{N} \sum_{j=1}^{N+1} d_{ij} y_{ij}\]

This is a linear programming problem, and furthermore, has the form of a transportation problem. It is especially easy to solve, because if one uses the column minima method of finding an initial solution, beginning with the first period, the solution so obtained will be optimal, and no iterations are required. The problem described here seems to have been developed independently by several individuals at roughly the same time. The transportation format for solving it was first published by Bowman.\(^{24}\). In actual use of this model, the problem would be re-solved at the beginning of every period using a planning horizon of \( N \) periods. However, whatever new information was
available would be included at the beginning of each period. Note that the model just considered is a periodic review model: i.e., decisions are made only at the beginning of each period.

Now a slightly different deterministic model will be studied in which it is assumed that production is carried out in lots rather than continuously. Here also the system will be a periodic review type. Again a planning horizon of $N$ periods is used, and the demand in each of these $N$ periods is predicted and then treated as deterministic. The ending inventory value is also specified. The problem is to determine in which periods to have a production run and what the size of the lot should be, if demands are to be met exactly in each period and the combined costs of setup and carrying inventory are to be minimized. This type of problem has been treated by Wagner and Whitin (25).

Let $A_j$ be the cost of a setup in period $j$ and $B_j$ the cost of carrying one unit in inventory for period $j$. This cost will be based on the inventory at the end of period $j$. It is assumed that the unit variable cost of producing the item is a constant, independent of the size of the production lot. The time required for production will here be ignored, although only the most trivial change in interpretation is required if there is a constant leadtime. The first assumption made is that there is no inventory on hand at the beginning of the first period. This is not unreasonable if the time considered to be the beginning of the first period can be set arbitrarily, since there would normally be no reason to have additional units arrive until those in inventory were used up.
Denote by $y_j$ the onhand inventory at the end of period $j$. Then it is desired to determine a set of $Q_j \geq 0$, $Q_j$ be the size of the production lot in period $j$ (this may be 0) which minimizes the cost $R$,

\begin{equation}
R = \sum_{j=1}^{n} \left[ A_j \delta_j + B_j y_j \right],
\end{equation}

where

\begin{equation}
y_j = y_{j-1} + Q_j - D_j, \quad j = 1, \ldots, N + 1; \quad y_0 = 0,
\end{equation}

and $D_j$ is the demand in period $j$. Furthermore,

\begin{equation}
\delta_j = \begin{cases} 
1 \text{ if } Q_j > 0 \\
0 \text{ if } Q_j = 0
\end{cases}.
\end{equation}

The formulae in Eq. 111 are simply material balance equations, and $y_{N+1}$ is the desired ending inventory.

This problem can easily be formulated in terms of the recurrence relations of dynamic programming. Let $\Lambda_n(\xi)$ be the minimum cost for the first $n$ periods when the onhand inventory at the end of period $n$ is $\xi$. Then

\begin{equation}
\Lambda_n(\xi) = \min_{Q_1, \ldots, Q_n} \left\{ \sum_{j=1}^{n} \left[ A_j \delta_j + B_j y_j \right] \right\},
\end{equation}

where Eq. 111 holds for $j = 1, \ldots, n - 2$, and

\begin{equation}
\xi = y_{n-1} + Q_n - D_n.
\end{equation}
Consequently,

\[ \Lambda_n(\xi) = \min_{Q_n} \left[ A_n \xi + B_n \xi + Q_1, \ldots, Q_n \min \left\{ \sum_{j=1}^{n-1} \left( A_j \xi + B_j \gamma_j \right) \right\} \right], \]

where

\[ \gamma_{n-1} = \xi - Q_n + D_n. \]

Therefore the recurrence relations of dynamic programming become

\[ \Lambda_n(\xi) = B_n \xi + \min_{Q_n} \left[ A_n \xi + \Lambda_{n-1} (\xi - Q_n + D_n) \right], n = 2, \ldots, N. \]

Then

\[ \min R = \Lambda_N(\gamma_N + 1). \]

The optimal solution could be obtained by solving sequentially the above recurrence relations. For this particular problem, however, a much simpler computational scheme can be devised.

To obtain this simplified computational scheme, first note that

\[ Q_j = 0 \text{ if } \gamma_{j-1} > 0; \text{ i.e., } Q_j \gamma_{j-1} = 0. \]

If this were not true, costs could be reduced by rescheduling \( \gamma_{j-1} \) units to period \( j \), thereby reducing inventory carrying charges. Therefore \( Q_j \) will always be equal to an integral number of periods' demand if \( Q_j > 0 \). It is also easy to show that if for \( \Lambda_k(0) \) it is optimal to have \( Q^*_k > 0 \), then \( Q^*_k > 0 \) for \( \Lambda_k(\xi), \xi > 0 \). An optimal policy for \( k \) periods, when nothing is onhand at the end of period \( k \), must then have the form that an order is produced at the beginning of period \( w \) which satisfies the demands in periods \( w \) through \( k \) and an optimal policy is followed in periods \( 1 \) through \( w-1 \), given that nothing is onhand at the end of period \( w-1 \). Thus
\[(118) \quad \Lambda_k(0) = \min_w \left\{ A_w + \sum_{j=w}^{k-1} B_j \left( \sum_{i=j+1}^{k} D_j \right) + \Lambda_{w-1}(0) \right\}, \]

or

\[(119) \quad \Lambda_k(0) = \min_w Y_k(w) \]

where

\[(120) \quad Y_k(w) = A_w + \sum_{j=w}^{k-1} B_j \left( \sum_{i=j+1}^{k} D_j \right) + \Lambda_{w-1}(0) \]

It can also be shown that if for \( \Lambda_{k-1}(0) \), the units which satisfy the demand in period \( k-1 \) are produced at the beginning of period \( v \), then in computing \( \Lambda_k(0) \), the units which satisfy the demand in period \( k \) will be produced at the beginning of period \( v \) or at a later period. In computing \( \Lambda_k(0) \), it is therefore unnecessary to consider the possibility that the units which satisfy the demand in period \( k \) were produced before \( v \). The resulting computational scheme, requiring only that Eq. 119 be determined, is very simple. Using this plan, one can work forward in time, progressively adding more periods but not being required to make changes in any periods preceding the last period, \( u \), for which \( Q_u^* > 0 \) in \( \Lambda_u(0) \).

The assumption of deterministic demand constitutes a limitation in the practical use of the two models developed in this section. Only infrequently does this assumption approximate reality closely enough to be useful. Furthermore, the type of cost structure and the assumptions made concerning the production process itself are not always sufficiently realistic.
STOCHASTIC DYNAMIC MODELS

When demands are assumed to be generated by a stochastic process, and when the mean rate of demand is changing, one immediately encounters the question of how the demand variable is going to be described mathematically. Usually, one considers only periodic review systems, treating as independent the random variables representing demand in different periods. The mean demand is permitted to shift from one period to the next. In reality, things are often far more complicated, since demands in different periods may be not independent but autocorrelated. None of the following models in which probability densities are explicitly introduced will attempt to account in any detail for such effects.

Let us consider first a dynamic N-period model. We will imagine that, if desired, an order can be placed at the beginning of each period. Let \( x_n \) be the demand in period \( n \); \( x_n \) will be treated as a discrete random variable, the density function of which is \( p_n(x_n) \). As before, attention will be limited to the backorders case. The inventory position at the beginning of the \( n \)th period, prior to the placing of any order, will be denoted by \( s_n \). The quantity ordered at the beginning of period \( n \) will be written \( Q_n (Q_n \geq 0) \). The procurement lead time will be assumed to be a constant \( \tau \). Let \( A_n \) be the fixed cost of placing an order in period \( n \) and let \( C_n(Q_n) \) be the cost of \( Q_n \) units in period \( n \). Here the unit cost of the item need not be a
constant. The expected costs of carrying inventory and of backorders in the period $t_n + \tau$ to $t_n + \tau + T$ will be denoted by $r_n(\xi_n + Q_n)$ if the inventory position at time $t_n$ (the beginning of period $n$) is $\xi_n + Q_n$ after any order is placed. Note that nothing can be done about the costs incurred between $t_1$ and $t_1 + \tau$, and that these costs have no influence on the decision as to how much to order at time $t_1$.

The minimum expected cost of operating the system for periods $n$ through $N$, when the inventory position of the system at the beginning of period $n$ prior to the placing of any order is $\xi$, will be denoted by $A_n(\xi)$. Then the $A_n(\xi)$ must satisfy the following recurrence relation

\begin{equation}
A_n(\xi) = \min_{Q_n} \left[ A_n \xi + C_n(Q_n) + r_n(\xi + Q_n) \right. \\
\left. + \sum_{x_n=0}^{\infty} \rho_n(x_n) A_{n+1}(\xi + Q_n - x_n) \right], \; n = 1, \ldots, N - 1
\end{equation}

and

\begin{equation}
A_N(\xi) = \min_{Q_N} \left[ A_N \xi + C_N(Q_N) + r_N(\xi + Q_N) \right].
\end{equation}

Then $Q_1^*$, the optimal quantity to order at the beginning of the first period, is determined by making this computation:

\begin{equation}
A_1(\xi_1) = \min_{Q_1} \left[ A_1 \xi_1 + C_1(Q_1) + r_1(\xi_1 + Q_1) \right. \\
\left. + \sum_{x_1=0}^{\infty} \rho_1(x_1) A_2(\xi_1 + Q_1 - x_1) \right].
\end{equation}
In order to perform this computation, $\Lambda_2(\xi)$ must be available which in turn means that $\Lambda_3(\xi)$, $\ldots$, $\Lambda_N(\xi)$ must have been determined previously. In certain situations, the system just considered will be stocking the item of interest for the indefinite future and the $N$ periods in the above model merely represent the planning horizon. At the beginning of each period, a computation would be made using a planning horizon of $N$ periods, in order to determine how much to order for the current period. Then $\Lambda_N(\xi)$ would be computed using Eq. 122. Note that no attempt is made to place any restriction on the ending inventory. If this is to be a valid approach, the planning horizon must be sufficiently long that the inventory at the end of the last period has little influence on the current decision.

In other situations, the item being studied will become obsolete in a relatively short time; then the above model will be applied over the remaining life of the item. Here, at the date of obsolescence, any items still in stock will be sold for their salvage value. In these situations, when one computes $\Lambda_N(\xi)$ in Eq. 122, the expected salvage gain should be deducted before minimizing over $Q_N$.

It should be noted that in the stochastic $N$ period model considered above, a numerical value can be found for $Q_1^*$. For the other $Q_j$, a function $Q_j^*(\xi)$ is obtained which indicates the quantity to order when at the time the order is being considered the inventory position is $\xi$. This is because one cannot decide how much to order at the beginning of period $j$ until the demands in periods $1$, $\ldots$, $j-1$ are known: i.e., until the beginning of period $j$ is reached.
For an N period dynamic model of the type just considered, a Kk policy will be optimal if the $r_j(y)$, $j = 1, \ldots, N$, are convex. This means that there exist numbers $K_j$ and $k_j$ such that, if at the beginning of period $j$, the inventory position is less than $k_j$, a sufficient quantity should be ordered to bring the level up to $K_j$. The values of $K_j$ and $k_j$ may, of course, vary with $j$. Proof of this can be found in Ref. 20.

The above model can be interpreted as a model dealing only with the problem of controlling inventory -- for spare parts in a military supply system, for example. It can also be interpreted as a combined production scheduling-inventory control model for the case in which the item is produced in lots rather than continuously. Then $A_j$ is the setup cost for period $j$, and $C_j(Q_j)$ is the variable cost of the $Q_j$ units (the unit cost may here vary with the lot size).

Next, we will consider two dynamic stochastic models for combined production scheduling-inventory control problems for the case where production of the item is continuous. As usual, however, it will be assumed that production decisions are made only periodically. It is possible to obtain models of widely varying degrees of complexity, depending on how much detail one wishes to include in describing the nature of the production process.

Consider first a model in which the only costs associated with the production of the units will be the unit variable cost of production, which is assumed to be independent of the production rate, and a cost for changing the production rate in a given period. At the beginning of each period, the production rate is fixed in advance. No change in the production rate is then considered until the begin-
ning of the next period. For problems of this sort, it is often convenient to imagine that all variables can be treated as continuous. A planning horizon of $N$ periods will be used. Let $f_j(x_j)$ be the probability density for the demand $x_j$ in period $j$, let $y_j$ be the net inventory at the beginning of period $j$, and let $w_j$ be the production rate (in units per period) for period $j$. When the production rate is changed at the beginning of a period, it will be assumed that both the rate at which units enter production and the rate at which they are finished are adjusted to the new value.

Thus the number of units produced in period $j$ will be $w_j$.

Denote by $r_j(y_j, w_j)$ the expected carrying costs and backorder costs for period $j$, when the initial net inventory is $y_j$ and the production rate is $w_j$. The cost incurred by any change in the production rate from the previous period will be denoted by $v_j(w_j, w_{j-1})$.

A possible form for the function $v_j$ might be the following:

$$
(124) \quad v_j(w_j, w_{j-1}) = \begin{cases} 
  a(w_j - w_{j-1}), & w_j - w_{j-1} > 0 \\
  b(w_{j-1} - w_j), & w_j - w_{j-1} < 0 
\end{cases} 
$$

Let $C_j$ be the unit variable production cost of the item in period $j$. The problem is to determine the values of the production rate for each of the periods in the planning horizon so as to minimize the expected costs of production, carrying inventory, and backorders.
The problem can be solved numerically with the aid of dynamic programming. Let \( \Lambda_n(\xi, \eta) \) be the minimum expected cost for periods \( n \) through \( N \) when the net inventory at the beginning of period \( n \) is \( \xi \), the production rate during period \( n-1 \) was \( \eta \), and an optimal policy is followed at period \( n \) and all future periods in setting the production rate. Then the recurrence relations take the form

\[
\Lambda_n(\xi, \eta) = \min_{w_n} \left[ C_n w_n + v_n(w_n, \eta) + r_n(\xi, w_n) \right. \\
+ \left. \int_{0}^{\infty} f_n(x_n)'(\xi + w_n - x_n, w_n) \, dx_n \right], \quad n = 1, \ldots, N - 1
\]

and

\[
\Lambda_N(\xi, \eta) = \min_{w_N} \left[ C_N w_N + v_N(w_N, \eta) + r_N(\xi, w_N) \right].
\]

Finally, if \( y_1 \) is the net inventory at the beginning of the first period, the minimum cost over the planning horizon is \( \Lambda_1(y_1, w_0) \). A numerical value can be found only for \( w_1 \), since the values of the other \( w_n \)'s cannot be determined until the demands in periods previous to the one for which \( w_n \) is to be determined are known. The problem would presumably be re-solved at the beginning of each period using a planning horizon of \( N \) periods. The present model differs mathematically in one very important way from the previous model for which the recurrence relations of dynamic programming are given by Eqs. 121 and 122. This difference lies in the fact that, in Eqs. 125 and 126, \( \Lambda_n \) is a function of two parameters whereas in Eqs. 121 and 122, only a single parameter was involved. Because of this, it can easily take a computer 100 times as long to solve Eqs. 125, 126 as it would to
solve Eqs. 121 and 122, since \( \Lambda_n \) must now be computed for all combinations of \([i, n]\). Furthermore, considerably more storage space will be required in the computer to store the tables of \( \Lambda_n \) because they now involve two arguments. No indication has been given above as to the method of computing \( r_j(y_j, v_j) \), the expected costs of carrying inventory and of stockouts for period \( j \). The computation is slightly different here than for lot production models, and the method for making it will now be illustrated in a special case. It now becomes necessary to introduce a distribution for the demand up to any time \( t \) in the interval. The unit of measure for time will be the length of period \( j \), so that \( t \) is the fraction of period \( j \) which has passed. Then let \( f_j(x, t) \) be the probability that the demand up to time \( t \) in period \( j \) lies between \( x \) and \( x + \text{dx} \). Units will be coming off the production line at a constant rate \( v_j \), so that the total production up to time \( t \) will be \( w_j t \). The expected carrying costs for period \( j \) are then

\[
I^C_j = \int_0^1 \int_0^{y_j + w_j t} \left( [y_j + w_j t - x] \right) f_j(x, t) \, dx \, dt
\]

(127)

\[
= I^C_j \left[ \frac{y_j + w_j t}{2} - \int_0^1 \mu(t) \, dt + B_j(y_j) \right].
\]

where \( I^C_j \) is the cost of carrying one unit for period \( j \), \( \mu(t) \) is the expected value of \( x \) for a given \( t \), and \( B_j(y_j) \) is the expected unit periods of backorders incurred during period \( j \).

Then

\[
B_j(y_j) = \int_0^1 \int_0^{y_j + w_j t} \left( [x - y_j - w_j t] \right) f_j(x, t) \, dx \, dt.
\]

(128)

Assume that there is a backorder cost for each unit backordered which
is proportional to the length of time for which the backorder exists, and let \( \hat{\Pi}_j \) be the cost of having a backorder on the books for period \( j \). The expected backorder costs for period \( j \) are then

\[
\hat{\Pi}_j B_j (v_j),
\]

and \( r_j(y_j, w_j) \) is the sum of Eqs. 127, 129. It is also possible to account for a fixed backorder cost, but this is a little more complicated and will not be considered here.

Beckmann(26) studied the nature of the optimal operating doctrine for the above model under the assumption that the mean rate of demand did not change with time, and that an infinite planning horizon was used. He showed that in this case the optimal operating doctrine has the following simple form. There exist two functions, \( w_1 = \phi_1(y) \), \( w_2 = \phi_2(y) \), such that if \( y \) is the onhand inventory at the beginning of a period, and if \( w \) was the production rate in the last period, then if \( \phi_1(y) \leq w \leq \phi_2(y) \), the production rate should not be changed; and if \( w < \phi_1(y) \), the production rate should be increased to \( \phi_1(y) \); and if \( w > \phi_2(y) \), the production rate should be reduced to \( \phi_2(y) \). The analysis he used, just as in the analysis referred to above in "Use of Functional Equations to Study Nature of Optimal Operating Procedures," depended critically on the convexity of the various cost components.

A model of the type just discussed would be difficult to use routinely in practice, except under highly unusual circumstances, because of the need for a very large computer and a considerable amount of computer time. Another model, developed by Holt, Muth, Modigliani, and Simon(7), which can be solved with little numerical computation will
now be considered. In many respects, the model is quite similar to
the one just discussed. The novel feature which introduces the
computational simplicity is based on the following result first
obtained and proved by Simon. If, in a dynamic programming problem
such as the one just discussed above, all the costs can be represented
as positive definite nonhomogeneous quadratic forms in the decision
variables, then, so far as the decision for the first period is con-
cerned, all random variables may be replaced by their expected values,
and the problem may be solved as if they were known with certainty.
In other words, assuming that the costs are positive definite quadratic
forms, if one minimizes the cost function obtained by replacing all
random variables with their expected values, the same values of the
decision variables will be obtained for the first period as those which
are obtained by minimizing the expected value of the cost. This allows
a tremendous reduction in computational effort, because the probability
distribution for the random variables can be eliminated. In fact,
simply by setting the appropriate partial derivatives to zero, one can
obtain equations which determine the optimal values of the decision
variables. This certainty equivalent for quadratic costs does not in
any way rely on the independence of demands in different periods. They
may have an arbitrary autocorrelation function. Thus, when the certainty
equivalent is used, the requirement that demands must be independent
in different periods may be dropped.

The Holt, Muth, Modigliani, and Simon model approximates all of
the costs by quadratic functions and therefore can use the certainty
equivalent to obtain the optimal decision for the first period. The
model is an aggregate one; if the factory is producing more than a single item, all of the items are aggregated together in making the decisions. Individual items are not considered separately. The basic decision variables which are to be determined in each period are the size of the work force and the aggregate level of production. These decisions then implicitly determine aggregate inventory levels as well. This model is somewhat more general than the model discussed above in that both the production for the period and the work force size are variables. In the above model, we assumed that setting the production rate determined the size of the work force. The present model is based on the assumption that for any given size of the work force there is a normal or standard aggregate production for the period. However, even though the work force is fixed, production can be increased over the normal value by use of overtime.

Consider then the details of the various costs involved. In the following, all C's will denote constants to be determined empirically, while $W_t$, $P_t$, $I_t$, and $S_t$ will denote respectively the number of men in the work force in period $t$, the aggregate production in period $t$, the net inventory at the end of period $t$, and the expected aggregate demand in period $t$. The cost components for period $t$ and their mathematical representation in the model are then as follows:

\begin{align*}
(130) \quad \text{Regular payroll cost} &= C_1 W_t + C_{01} \\
(131) \quad \text{Cost of hiring and layoffs} &= C_2 (W_t - W_{t-1} - C_{02})^2 \\
(132) \quad \text{Cost of overtime} &= C_3 (P_t - C_4 W_t)^2 + C_5 P_t - C_6 W_t + C_7 P_t W_t \\
(133) \quad \text{Inventory backorder, and setup costs} &= C_8 \left[ I_t - (C_{03} + C_9 S_t) \right]^2
\end{align*}
The way in which the C's are determined results in the cost terms being positive definite nonhomogeneous quadratic forms in the decision variables. The total cost for \( N \) periods is then

\[
R = \sum_{t=1}^{N} \left( c_1 W_t + c_{01} + c_2 (W_t - W_{t-1} - c_{02})^2 + c_3 (P_t - c_{03})^2 \right)
\]

\[(134)\]

\[+ c_5 P_t - c_6 W_t + c_7 P_t W_t + c_8 \left[ I_t - (c_{03} + c_9 S_t)^2 \right].\]

This cost expression is the certainty equivalent of the expected cost for \( N \) periods since the random variables representing the demands in the different periods were replaced by their expected values \( S_t \). It is Eq. 134 that is minimized to obtain the optimal values of \( W_t \) and \( P_t \). Note that the problem is re-solved at the beginning of each period and only \( W_t \) and \( P_t \) are computed each time. The ending inventories \( I_t \) appearing in Eq. 134 must satisfy the material balance equations.

\[
(135) \quad I_t = I_{t-1} + P_t - S_t.
\]

The variables \( W_t, P_t \) must also be non-negative, but since in practice they always turn out to be so, it is unnecessary to include the non-negativity constraints explicitly.

To determine the optimal values of the decision variables, the partial derivatives of \( R \) with respect to these variables are set to zero, and the resulting set of equations solved for the decision variables. Since \( R \) is quadratic, the set of equations will be linear, and, consequently, the optimal values of \( W_t \) and \( P_t \) will depend linearly on the values of \( W_0, I_0, \) and the \( S_t \).
\[
F_1^* = \alpha_0 + \sum_{t=1}^{N} \alpha_t S_t + \delta_0 W_0 + \delta_1 I_0
\]

\[
W_1^* = \beta_0 + \sum_{t=1}^{N} \beta_t S_t + \gamma_0 W_0 + \delta_1 I_0
\]

In actuality, Holt, et al., found it easier to use an infinite planning horizon and to solve the resulting infinite set of equations by a special technique that will not be considered here. Fortunately, the weights \(\alpha_t\) and \(\beta_t\) decrease very rapidly as \(t\) becomes large, and therefore a reasonably short planning horizon is usually adequate in practice.

The above model has actually been implemented in several companies, and the authors claim that it has led to an improvement in performance. They concede, however, that it is very difficult to make realistic evaluations as to what sort of improvement there was. In the real world, of course, costs cannot be represented by strictly quadratic functions. In this event, the nature (i.e., the shape) of the probability distribution for demands does have an influence on the optimal decisions. No attempts have been made to determine how important these effects can be. The only claims made by the authors are that the model (often referred to as a "linear decision rule model") has yielded better results than were obtained with the methods previously in use at the companies where the model has been implemented.

**THE PREDICTION PROBLEM**

All of the dynamic models considered above have assumed that the mean demand for each period in the planning horizon, and possibly the probability distribution for demand in each period, were known. In
practice, it can be a very difficult task to estimate even the expected
demand for each future period, not to mention the distribution of
demand for each period.

The methods which can be used to predict the demand in future
periods vary widely with the circumstances. Sometimes detailed
economic forecasts are warranted. In other instances nothing more
than historical data will be used.

An especially difficult prediction problem (and also one that
is of special interest in the military) is predicting demands for
spare parts. The procedure which has been used in the military is
to establish a usage rate on the basis of historical data. This rate
gives the average number of spares needed per hour of usage for one
system. Next, the total number of hours of usage for all systems is
predicted over the period of interest. This is multiplied by the
usage rate and the resulting number is used for the expected spares
demand over the relevant time interval. Note that with this procedure,
no probabilistic considerations enter in. Aside from random fluctuations,
predictions can be in error both because the usage rate was estimated
incorrectly, and because the predicted total hours of usage was estimated
incorrectly. It is exceptionally difficult to obtain accurate values
of the usage rate for low demand spare parts since the system becomes
obsolete before a long enough demand history is available to make an
accurate estimate. A great deal of work has been expended in trying to
find ways to predict usage rates accurately, but, so far, these attempts
have met with little success.

Consider next the techniques for making predictions in the case
where only historical data are used, and where there are no strong
seasonal patterns. The two most widely employed procedures are fitting a least squares line to the historical data, and exponential smoothing. These techniques will be considered briefly. Let us imagine that the prediction techniques are going to be used in a periodic review system, and that a new prediction will be made at the beginning of each period.

If the demand $D_j$ in period $j$ is predicted using the linear relation $D_j' = a_j + b$, where $a$, $b$ are constants determined by minimizing

$$ F = \sum_j (D_j - D_j')^2 = \sum_j (D_j - a_j - b)^2, $$

and $D_j'$ is the predicted demand for period $j$, then the method is called "least squares" prediction using a linear prediction relation. Imagine that $t$ is the time at which the prediction is being made, and one desires to predict the demand in period $j$ which covers the time interval $t + (j-1)T$ to $t + jT$ using the demands in the $N$ periods $0, -1, \ldots, -(N-1)$, where period 0 extends from $t$ to $t-T$, etc. To determine the values of $a$, $b$ which minimizes Eq. 138, the equations $\frac{\partial F}{\partial a} = \frac{\partial F}{\partial b} = 0$ are solved. This yields for $D_j'$

$$ D_j' = \frac{X}{N} + 12 \left[ \frac{Y + \frac{N-1}{2} X}{N(N-1)} \right] \left[ j + \frac{N-1}{2} \right] $$(139)

where

$$ X = \sum_{j=0}^{-(N-1)} D_j, \quad Y = \sum_{j=0}^{-(N-1)} j D_j. $$

Only $X$ and $Y$ need to be computed to obtain new estimates $D_j'$ each time the prediction is repeated. These need not be recomputed from their definitions, but may instead be computed using the recurrence relations
\[(141) \quad \hat{x} = D_1 + x - D_{-(N-1)}; \quad \hat{y} = y - x + ND_{-(N-1)}, \]

where \(\hat{x}\) and \(\hat{y}\) are the values of \(x\) and \(y\) using the data for periods 1, 0, ..., \(-(N-2)\).

Note that in Eq. 139, \(\frac{x}{N}\) is the average demand for the past \(N\) periods. Thus the predicted demand for period \(j\) is the average demand over the past \(N\) periods plus a trend correction. The uncertainty in the estimate of \(D_j\) increases with \(j\). If sufficient historical data are available, one can obtain a histogram of the forecast errors (these representing the difference between the actual demand and the predicted demand) by using Eq. 139 to predict the demand in future periods for each period over which data are available, and then comparing the predicted value with the actual value. In a dynamic model, it would really be the appropriate distribution of forecast errors that one would want to use whenever a distribution is needed. Furthermore, a way to select the value of \(N\) appearing in Eq. 139 is to select that \(N\) which minimizes the variance of the distribution of forecast errors. Unfortunately, it is often true in practice that not enough of a demand history is available to yield a detailed histogram for the forecast errors. Part of the reason for this is that the processes generating the demands are not stationary.

The least squares technique suffers from the difficulty that the demands for the past \(N\) periods must always be available. This is not a great problem for a small number of items, but when the number increases substantially and the computations are made on a digital computer, a lot of storage space is required for this information. The exponential smoothing technique often reduces considerably the amount of past data
which must be kept available. Given a discrete time series of data
with \( f_j \) being the value of the variable for period \( j \), the exponentially
smoothed value for period \( j \), denoted by \( \bar{f}_j \), is defined in terms of \( f_j \)
and \( \bar{f}_{j-1} \) as

\[
\bar{f}_j = \gamma f_j + (1 - \gamma) \bar{f}_{j-1}, \quad 0 < \gamma < 1;
\]

i.e., the smoothed value for period \( j \) is \( \gamma \) times the value of the
variable for period \( j \), plus \( 1 - \gamma \) times the smoothed value for period
\( j-1 \). The value of \( \gamma \) indicates how much weight is given to the current
value of the variable in carrying out the smoothing. Repeated
substitution reveals that all previous values are used in computing
the smoothed value for period \( j \), i.e.,

\[
\bar{f}_j = \gamma \sum_{i=0}^{\infty} (1 - \gamma)^i f_{j-i}.
\]

To estimate the demand in period \( j \) using exponential smoothing,
one might try

\[
D'_j = \bar{D}_0.
\]

This would be equivalent to using \( D'_j = \frac{x_j}{N} \) in the least squares case,
and it does not include a trend correction. To determine the correction
needed to account for a linear trend, assume that demand has been
increasing at a constant rate of \( \delta \) units per period for all past time,
so that the demand in period \( j \) can be written \( \alpha + \delta j \). Then

\[
\bar{D}_0 = \gamma \alpha + \gamma (1-\gamma) [\alpha - \delta] + \gamma (1-\gamma)^2 [\alpha - 2\delta] + \gamma (1-\gamma)^3 [\alpha - 3\delta] + \ldots
\]

\[
D'_0 = \alpha - \frac{1-\gamma}{\gamma} \delta,
\]
whereas in actuality $D_j = c + 8j$. To obtain $D_j$ from $\bar{D}_o$, it is necessary to add

$$\left[ j + \frac{1 - \gamma}{\gamma} \right] 8.'$$

In practice, $8$ must also be estimated from historical data.

Exponential smoothing can be used for this purpose, too. Let

$$(146) \quad 8_j = \bar{D}_j - \bar{D}_{j-1}.$$  

The smoothed value of $8$ will be then taken to be

$$(147) \quad \bar{8}_o = \gamma 8_o + (1 - \gamma) \bar{8}_{o-1}.$$  

The predicted value for the demand in period $j$ is then taken to be

$$(148) \quad D'_j = \bar{D}_o + \left[ j + \frac{1 - \gamma}{\gamma} \right] 8_o.$$  

With exponential smoothing it is only necessary to keep in storage $\bar{D}_{o-1}$ and $8_{o-1}$ in addition to the past period's demand. When possible, $\gamma$ may be chosen to minimize the variance of the distribution of forecast errors.

The above technique is called exponential smoothing for the following reason. If $\gamma = \lambda \Delta t$ in Eq. 142, and the limit is taken as $\Delta t = 0$ while holding $\lambda$ constant, the following differential equation is obtained:

$$(149) \quad \frac{1}{\lambda} \frac{d\bar{f}}{dt} + \bar{f} = f(t).$$

The solution to this equation is

$$(150) \quad \bar{f} = \lambda \int_{-\infty}^{t} e^{-\lambda(t-\zeta)} f(\zeta) d\zeta.$$
if $\bar{f}(\infty) = 0$. Thus $\bar{f}$ is an exponentially weighted sum of all previous values of $f(t)$.

When a seasonal pattern is the most important aspect of the demand pattern, one will frequently predict that sales over a given interval in the current season will be equal to those last season for the corresponding interval, plus a percentage correction appropriate to the general state of business. It is not too easy to obtain a distribution of forecast errors in this case, since the errors may depend on the part of the season under consideration.

**REDISTRIBUTION MODELS**

The inventory models discussed thus far have implicitly assumed that stocks are located at a single installation. If several installations are involved, such as several depots in logistics problems, it may be inappropriate to attempt to utilize inventory models without simultaneously considering the effect of transportation costs on the logistics system. The costs involved in redistributing stocks among the various depots can be of considerable magnitude. In fact, even for single installation models, there exists a trade-off between the procurement leadtime and inventory levels, and the optimizations discussed above are valid only when optimal leadtime is selected.

Redistribution models developed for the Navy's Bureau of Supplies and Accounts have included the costs of redistribution along with inventory costs and have provided decision rules for procurement, allocation, and redistribution.\(^{(27)}\) The potentially large effect that redistribution costs exert both on the reorder point for the system and on the quantity procured was pointed out. A detailed description of this model is beyond the scope of this report.
MULTI-ECHELON MODELS

Relatively little analytical work has been done on the multi-echelon systems. A single exception is some work done by Scarf and Clark \(^{28}\)(29) for a very special structure under rather restrictive assumptions. We will now describe their model briefly. They imagine the multi-echelon structure to consist of a one-dimensional chain of stockage points such that each stockage point ships only to the stockage point at the next lowest level. All redistribution problems are eliminated since there is only a single stockage point at each echelon level. Furthermore, it is impossible to by-pass and ship, say, to a stockage point two echelons down the chain. In addition, the following assumptions are made:

1. The system operates using a periodic review procedure and the time between reviews is the same for each echelon.

2. Demands originate only at the lowest echelon and nowhere else. The demand rate is allowed to vary with time but only a finite planning horizon is considered.

3. There is no fixed cost of placing an order at any but the highest echelon. At the highest echelon, a fixed cost is allowed.

4. Each echelon backorders demands when out of stock.

5. At the lowest level, a backorder cost is incurred which is proportional to the number of backorders existing at the end of the period. This assumption is not essential. Backorders can be costed in the more general ways discussed earlier.

6. Leadtimes are constant and multiples of the period length.

7. Carrying costs are based on the onhand stock at the beginning of a period. This assumption is not essential either, but can be
generalized to other costing procedures.

At any point in time, the echelon stock for a given echelon is defined to include the stock on hand at that echelon plus the stock on hand at all the lower echelons plus any stock in transit to lower echelons. Echelon stock, then, is the total amount of stock in the system at the highest and any lower echelons.

Note that the only point at which the inventory carrying charge is important is in deciding how much to bring into the system. It is irrelevant in deciding how to move units around within the system unless there is a differential carrying charge at each echelon. Thus, for each period, one can base the inventory carrying charges on the echelon stock, and for all echelons except the highest, this carrying charge is simply the differential charge between the echelon under consideration and the one immediately above it. Stock in transit is charged at the rate corresponding to the echelon from which it was shipped.

Now it can be demonstrated that with the above assumptions the multi-echelon system can be solved sequentially one echelon at a time using the dynamic programming approach presented earlier. Since there are no fixed costs of placing an order, except at the highest echelon, it is clear that all echelons except that one will follow a policy of bring their echelon inventory position up to a level $K_{jn}$. The level $K_{jn}$ depends on the echelon $j$ and time period $n$. The computational procedure is as follows: Begin with the lowest echelon, call it 1, and compute $K_{in}$ $n = 1, \ldots, N$ (assume that there are $N$ time periods in all). The expenses involved will be transportation costs (proportional to the number of units sent), differential holding costs, and backorder
costs. In the computation, assume that any order placed will be supplied by the next higher echelon (even though it may not be). Now move to the next highest echelon, calling it 2. Repeat the computation and determine the $K_{2n}$. The costs will again consist of transportation costs, differential carrying costs, and shortage costs. The important thing here is to specify the shortage cost for the second echelon. This cost is the expected increment in the cost at Echelon 1 because the stock at Echelon 2 is insufficient to bring Echelon 1 up to $K_{1n}$. The computation of this cost involves the numbers $K_{1n}$ and hence it is not possible to optimize Echelon 2 until Echelon 1 has been optimized. The same procedure is repeated at higher echelons. Scarf and Clark (28) prove that the computational procedure described yields an optimal solution. A proof will not be repeated here because of its length.

DYNAMIC STABILITY AND FREQUENCY RESPONSE OF COMPLEX SYSTEMS - DETERMINISTIC INPUTS

In studying complex inventory supply systems, one is often much more interested in the dynamic stability and frequency response of the system than in an optimal operating doctrine derived under the assumption that the stochastic processes associated with the system are not changing with time. There are a variety of reasons for this. First of all, it is well known that complex feedback systems may be dynamically unstable, and an unstable system will typically be very difficult and unusually expensive to operate effectively. Then there are often seasonal or longer range periodic variations in the demand pattern in the real world. It is quite possible that a system will amplify these variations
and change the phase. It is important, therefore, to know how the system responds to a sinusoidal input of various frequencies. It is also true that random fluctuations (i.e., noise) in the input to the system may actually lead to unstable oscillations in part of the system. Finally, it may not be possible to specify an optimal operating doctrine for all stocking points in the system since different organizations control the different stocking points, and the best that can be done is to try to compensate for the others' behavior in order to avoid certain undesirable characteristic modes of operation.

Most of the analytical work in this area has been done in electrical engineering in the course of studying electromechanical servomechanisms. Little has been done to apply the results to inventory systems or to modify the analytical results already available to make them more useful in the analyses of inventory problems.

The usual approach to the analysis of the dynamic behavior of some system is to formulate a set of differential or difference equations which represent the way the system behaves, and to study the solutions to these equations. Normally, it is assumed the system can be satisfactorily represented as a linear one so that only linear differential or difference equations appear in the mathematical model. This is done because it is much easier to obtain analytical solutions to linear equations and, in addition, systems can often be adequately represented, over a limited operating range, as being linear.

Consider now a system in which there is just a single input (perhaps consumer demand) and some system variable \( y \) that is of interest (\( y \) might be the factory production rate). If the variable \( y \) is a continuous function of the time \( t \), then the assumption normally made in systems analysis work is that \( y(t) \) is related to the input \( x(t) \) by
a linear differential equation with constant coefficients of the form

\[ \sum_{j=0}^{n} a_j \frac{d^j y}{dt^j} = \sum_{u=0}^{m} b_u \frac{d^u x}{dt^u} . \]

We may consider a factory warehouse as a simple example of how such a differential equation can arise. Let \( x(t) \) be the rate at which orders for units are received from wholesalers at the warehouse. Suppose that the following sort of rule is used to determine the rate \( y \) at which the warehouse orders units from the factory

\[ y = x + \alpha (I_d - I_a) , \]

where \( I_d \) is the desired inventory level in the factory warehouse and \( I_a \) is the actual onhand inventory level. Imagine that

\[ I_d = kx , \]

so that the desired inventory is a constant times the current rate of demand, i.e., the warehouse wants a fixed number of weeks supply on hand.

Let \( r \) represent the rate at which units are received from the factory at the warehouse; then if it is assumed that the shipping rate is equal to the demand rate,

\[ \frac{dI_a}{dt} = r - x . \]

It remains to relate \( r \) and \( y \). Usually, there will be a delay before orders appear as finished goods. The simplest relation with no delay would be \( r = y \), so that the ordering rate and rate of receipt of units
were equal. Another possibility would be to introduce a fixed delay, so that \( r(t) = y(t - \tau) \). Differential equations in which the variables are not all evaluated at the same point in time tend to be difficult to handle; hence the discrete delay causes problems. A more useful concept is the continuous delay in which the output rate is a weighted sum of the input rates over all previous times, i.e.,

\[
(155) \quad r = \int_{-\infty}^{t} G(t, \zeta) y(\zeta) \, d\zeta ,
\]

and \( G(t, \zeta) \) is called the Green's function for the delay. One case that is especially easy to handle is that when

\[
(156) \quad G(t, \zeta) = \mu e^{-\mu(t-\zeta)} .
\]

Then the delay is called a first order exponential delay, and

\[
(157) \quad r = \mu \int_{-\infty}^{t} e^{-\mu(t-\zeta)} y(\zeta) \, d\zeta .
\]

If Eq. 157 is differentiated with respect to \( t \), one obtains

\[
\frac{dr}{dt} = -\mu^2 \int_{-\infty}^{t} e^{-\mu(t-\zeta)} y(\zeta) \, d\zeta + \mu y = -\mu r + \mu y ,
\]

or

\[
(158) \quad \frac{1}{\mu} \frac{dr}{dt} + r = y .
\]

Equation 158 explains why the first order exponential delay is so useful -- the output is related to the input by a first order differential equation, and the delay can be represented nicely within the desired linear differential equation framework. It is not hard to show that the average delay is \( 1/\mu \). A first order delay can be represented
symbolically as

\[ \left( \frac{1}{\mu} s + 1 \right)^n r = y \]

where \( s = \frac{d}{dt} \). It is possible to define mathematically higher order exponential delays by connecting in series a sequence of first order delays. Thus the differential equation relation the output rate to the input rate for an nth order delay is

(159) \[ \left( \frac{1}{\mu} s + 1 \right)^n r = y \]

Note that if in Eq. 155 \( y \) is taken to be the Dirac delta function \( \delta(0) \), then \( r = G(t,0) \), and hence, the Green's function gives the output rate of the delay as a function of time for a unit impulse input at time zero. The shapes of the Green's functions for various orders of exponential delays are shown in Fig. 9.

Fig. 9 — Shapes of the Green's functions for various orders of exponential delays
After this digression on the representation of delays, the example will be again considered. Suppose that \( r \) is related to \( y \) by a first order delay, i.e., that Eq. 158 holds. It is now possible to derive a single equation relating \( y \) and \( x \). Differentiating Eq. 152 and using Eqs. 153 and 154, one obtains

\[
\frac{dy}{dt} = (1 + ky) \frac{dx}{dt} - \alpha(r - x).
\]

Differentiating again, one obtains

\[
\frac{d^2y}{dt^2} = (1 + ky) \frac{d^2x}{dt^2} - \alpha(\frac{dr}{dt} - \frac{dx}{dt}).
\]

Multiplying Eq. 161 by \( 1/\mu \) and adding to Eq. 160, one obtains after using Eq. 158

\[
\frac{d^2y}{dt^2} + \mu \frac{dy}{dt} + \mu \omega^2 = (1 + ky) \frac{d^2x}{dt^2} + [\mu + \alpha(\mu + 1)] \frac{dx}{dt} + \mu \omega x.
\]

Equation 162 is the differential equation relating the rate at which orders are sent to the factory to the rate at which orders are received at the warehouse. Note that it is a linear differential equation of second order. The dynamic behavior of the system is completely described by Eq. 162. The equation is linear because, in the derivation, negative ordering rate, negative inventories, etc., were allowed -- no nonlinear cutoffs were introduced. Therefore, such an equation might be expected to be valid only when deviations from some standard operating point were not too large. One could make the representation of the system much more realistic. The above example merely illustrates the procedure to be used. However, as the realism of representation is increased, the order of the equation will also usually increase quite rapidly.
In order to study the dynamic behavior of systems like the above, it is necessary to examine the nature of the solutions to differential equations such as

\[ \sum_{j=1}^{n} a_j \frac{d^j y}{dt^j} = f(t), \quad a_n \neq 0. \]  

(163)

Let \( y_p(t) \) be any solution to Eq. 163, and \( y(t) \) any other solution to the same equation. Then \( y_h(t) = y(t) - y_p(t) \) must satisfy

\[ \sum_{j=1}^{n} a_j \frac{d^j y_h}{dt^j} = 0. \]  

(164)

This is called the homogeneous equation. The general solution to Eq. 163 can be written as the sum of a particular solution \( y_p(t) \) and a solution \( y_h(t) \) to the homogeneous equation. Note that if \( g_1(t) \), \( g_2(t) \) are solutions to the homogeneous equation, then so is \( \alpha_1 g_1(t) + \alpha_2 g_2(t) \) for any constants \( \alpha_1, \alpha_2 \). There cannot be more than \( n \) linearly independent solutions to Eq. 164, since the Wronskian for every set of \( n + 1 \) solutions vanishes identically. Furthermore, there are always at least \( n \) linearly independent solutions (these will be exhibited). Thus there are always precisely \( n \) linearly independent solutions to the homogeneous equation; hence the general solution can be written

\[ y(t) = y_p(t) + \sum_{j=1}^{n} \alpha_j g_j(t), \]  

(165)

where the \( \alpha_j \) are arbitrary constants.

To determine the \( g_j(t) \), suppose that a solution of the form \( g(t) = e^{rt} \) is tried. Substitution of this into Eq. 164 shows that \( g(t) \) will be a solution if

\[ \sum_{j=1}^{n} a_j r^j = 0. \]  

(166)
This is called the characteristic equation for the differential equation. It will always have n roots if a root is counted a number of times equal to its multiplicity. The roots may be real or complex. If all n roots are different, the n solutions \( g_j(t) = e^{\gamma_j t} \) are linearly independent (since the Wronskian does not vanish) and all n linearly independent solutions have been found. If the root \( \gamma_k \) has multiplicity \( m \geq 2 \), then it can be shown that \( e^{\gamma_k t}, te^{\gamma_k t}, \ldots, t^{m-1} e^{\gamma_k t} \) are linearly independent solutions to the homogeneous equation. Thus, in every case, a set of n linearly independent solutions to the homogeneous equation have been found. When a root is complex, i.e., \( \gamma = a + bi \), then its complex conjugate \( a - bi \) will also be a root. Thus

\[
(167) \quad e^{at} [\cos bt + i \sin bt], \quad e^{at} [\cos bt - i \sin bt]
\]

are solutions to the homogeneous equation. Equally useful are two linearly independent solutions to the homogeneous equation which it is often convenient to use instead of those given in Eq. 167:

\[
(168) \quad e^{at} \cos bt, \quad e^{at} \sin bt
\]

A system of the type being considered is said to be dynamically stable if a bounded input results in a bounded output, and if when the system, being initially at rest, is shocked, it will ultimately return to rest again. The Dirac delta function is the activating shock. The solution to Eq. 163 for a delta function input at time 0 will now be determined. If the system was initially at rest, then \( \frac{d^j y}{dt^j} = 0, \quad j = 0, 1, \ldots, \) for \( t < 0 \). Furthermore, Eq. 164 is satisfied for \( t > 0 \), i.e., the solution has the form
\[ y = \sum_{j=1}^{n} \alpha_j g_j(t) \]

for \( t > 0 \). And

\[ \int_{0^-}^{0^+} \left\{ \sum_{j=1}^{n} a_j \frac{d^j y}{dt^j} \right\} dt = \int_{0^-}^{0^+} \delta(t) dt = 1 \]

Now everything except the highest order derivative must be continuous at \( t = 0 \) (because any jumps would lead to delta function discontinuities in the higher derivatives, and this would not fit in with the differential equation). Thus Eq. 170 reduces to

\[ a_n \int_{0^-}^{0^+} \frac{d^n y}{dt^n} = a_n \left[ \left( \frac{d^{n-1} y}{dt^{n-1}} \right)_{0^+} - \left( \frac{d^{n-1} y}{dt^{n-1}} \right)_{0^-} \right] = a_n \left( \frac{d^{n-1} y}{dt^{n-1}} \right)_{0^-} = 1. \]

It must also be true, of course, that

\[ \left( \frac{d^j y}{dt^j} \right)_{0^-} = 0, \quad j = 0, 1, \ldots, n-2. \]

Thus \( n \) equations are obtained to determine the \( n \) values of \( \alpha_j \).

They are

\[ \sum_{j=1}^{n} \alpha_j \left( \frac{d^k g_j}{dt^k} \right)_{0^-} = 0, \quad k = 0, 1, \ldots, n-2, \]

\[ \sum_{j=1}^{n} \alpha_j \left( \frac{d^{n-1} g_j}{dt^{n-1}} \right)_{0^+} = \frac{1}{a_n} \]

The matrix of coefficients is simply the Wronskian matrix, which is nonsingular. Thus, there is a unique solution for the \( \alpha_j \), and in general, each \( \alpha_j \neq 0 \).

Now in order that the system be stable, it must be true that \( y \) given by Eq. 169 must satisfy
\[
l_{t \to \infty} y(t) = \sum_{j=1}^{n} \alpha_j \left[ \lim_{t \to \infty} g_j(t) \right] = 0,
\]

and, in general, this will be true if and only if

\[
\lim_{t \to \infty} g_j(t) = 0.
\]

Equation 175 will hold, however, only if the real part of \( \gamma_j \) is negative. Thus one obtains the important criterion for dynamic stability of the system that, in general, the system will be stable if and only if all roots of Eq. 166 lie in the left half plane. The behavior of the system for various positions of the roots in the complex plane is shown in Fig. 10.

![Graph showing the effect of root position on dynamic stability](image)

**Fig. 10 — Effect of root position upon dynamic stability of the system**

Note that the stability criterion says nothing about how stable the system is, i.e., how long it takes to settle down after being shocked. This depends on how far the roots are to the left of the imaginary
axis. Those parts of any stable solution to Eq. 163 which die out with time are called transient components. A system is unstable when the transients do not die out. The transients are those parts of the solution to Eq. 163 which came from solving the homogeneous equation. It should be observed that even a stable system may exhibit an undesirable oscillatory behavior when shocked.

One of the major analyses carried out in the design of any feedback control system in electrical engineering is a stability analysis. The parameters of the system must be chosen in such a way that the system is stable and any oscillatory behavior is within tolerable limits. The inventory-production scheduling example given earlier for which the differential equation is Eq. 162, lends itself easily to a stability analysis. The system is always stable, but will exhibit oscillatory behavior when shocked if \( \mu < \frac{4\alpha}{\omega} \). This places a restriction on the parameters \( \mu \) and \( \alpha \). A graphical interpretation of the behavior of a system as it changes from an unstable to a stable non-oscillatory mode of operation will be presented in the portion of this Memorandum which deals with simulation.

In general, it is very difficult to determine explicitly the roots of the characteristic equation when \( n \geq 3 \). A number of techniques have been developed in electrical engineering for analyzing the stability of a system without actually finding the roots of the characteristic equation. Perhaps the best known of these is called the Nyquist criterion. It makes use of a theorem from function theory on the number of zeros and poles inside a closed contour. The method will not be discussed here because it is really not especially helpful in studying complex inventory control systems. In most systems employed by electrical engineering, the order of the characteristic equation is fairly low -- say, six or less -- whereas for complex inventory systems the order of
the equation can easily be 20 or more. It seems to be impossibly laborious to apply the methods developed in electrical engineering to such cases. Very little work has been done in studying whether the techniques of electrical engineering can be extended to handle the sorts of equations obtained in studying inventory systems, or whether the representations of inventory systems can be simplified sufficiently to use the existing electrical engineering techniques.

In the above it was assumed that the input and output variables were continuous and differentiable functions of time (except at isolated points). For some systems, the input and output are defined only at discrete, equally spaced intervals of time, so that the input can be written \( x(k\Delta t) = x_k, \ k = 0, 1, \ldots, \) and the output can be written \( y(k\Delta t) = y_k, \ k = 0, 1, \ldots. \) Then, in place of the differential Eq. 151, one will have an nth order difference equation

\[
\sum_{j=0}^{n} a_j y_{k+j} = \sum_{u=0}^{m} b_u x_{k+u}
\]

(Solution (176))

Solutions to this equation have the form \( y^k \) where \( y \) satisfies Eq. 166. Such systems are called sampled data systems. A sample data system will be stable if and only if the roots of Eq. 166 lie inside the unit circle in the complex plane.

Another interesting property of a control system is its frequency response, i.e., how the phase and amplitude of a sinusoidal input is modified as it passes through the system. To study this question, imagine that the output of the system is related to the input by Eq. 151. Instead of using a sine or cosine function as an input, it is convenient to use the complex form \( x = \rho e^{j\omega t}. \) Suppose that an attempt is made to
find a solution for the output of the form \(y = \rho_2 e^{i(\omega t + \phi)}\), i.e., a sinusoidal output with the same frequency but a possibly different amplitude and phase angle. Upon substitution of these into Eq. 151 one finds that \(\rho_2, \phi\) must satisfy

\[
\rho_2 e^{i\phi} \sum_{j=0}^{n} a_j (i\omega)^j = \rho_1 \sum_{u=0}^{m} b_u (i\omega)^u,
\]

or

\[
\frac{\rho_2}{\rho_1} = \left| \sum_{u=0}^{m} b_u (i\omega)^u \right| \left| \sum_{j=0}^{n} a_j (i\omega)^j \right|^{-1}
\]

(177)

\[
\tan \phi = \frac{\text{Im} Z}{\text{Re} Z}, \quad Z = \sum_{u=0}^{m} b_u (i\omega)^u \sum_{j=0}^{n} a_j (i\omega)^j
\]

(178)

Thus it is possible to determine values of \(\rho_2\) and \(\phi\) such that \(y\) is a solution. Hence the output will be sinusoidal with the same frequency as the input, with the amplification and phase angle being given by Eqs. 177 and 178 respectively. Note that both the amplification and the phase angle may depend on the frequency \(\omega\). Plots of \(\rho_2/\rho_1\) and \(\phi\) as a function of \(\omega\) might look something like those shown in Figs. 11, 12. A knowledge of what these curves are like is important in the design of electro-mechanical systems and in the analysis of inventory systems as well.

In attempting to solve equations such as Eq. 151, the use of Laplace transforms is especially helpful because it enables one to account for the initial conditions more conveniently than when one attempts to determine the \(\alpha_j\) directly. The Laplace transform of a
function \( f(t) \) is defined to be

\[
F(s) = \int_0^\infty e^{-st} f(t) \, dt
\]

and \( F \) is to be interpreted as a function of the complex variable \( s \).

A function is uniquely determined by its Laplace transform: i.e., the inverse transform is unique. Note that the Laplace transform of \( \frac{d^j F}{dt^j} \) is

\[
s^j F(s) = \sum_{u=0}^{j-1} s^{j-u-1} \left( \frac{d^u f}{dt^u} \right)
\]

Consequently, if \( Y(s) \), \( X(s) \) denote the Laplace transforms of \( y \), \( x \) respectively, we see, upon taking the Laplace transform of Eq. 151 that for a system initially at rest

\[
Y(s) = H(s) \, X(s),
\]
where

$$H(s) = \sum_{u=0}^{m} b_u s^u \sum_{j=0}^{n} a_j s^j$$

(182)

$H(s)$ is referred to as the transfer function for the system, and in cases of practical interest $m < n$, so that $H(s)$ is a meromorphic function whose poles lie at the zeros of the characteristic equation. If $h(t)$ is the inverse Laplace transform of $H(s)$, it is easy to show that

$$y(t) = \int_{0}^{t} h(t-\zeta) x(\zeta) d\zeta,$$

(183)

so that $y(t)$ is the convolution of $h(t)$ and $x(t)$. Note that, in terms of the transfer function, Eq. 177 becomes $p_2/p_1 = |H(i\omega)|$.

Up to this point, we have studied the means for analyzing the stability and the frequency response of a control system. The third important characteristic of such a system is its ability to track various sorts of signals. This tracking ability is usually measured by two different concepts. First, if the input signal is suddenly changed from one form to another, there is the question of how long it will take the system to settle down after the change. Then, there is the question of whether or not there will be any steady state error. Control systems usually operate by trying to reduce the value of some variable to a desired value (in the example above, we have tried to keep the actual inventory level equal to the desired inventory level). It is well known that no system can be designed to have a zero steady state error for all inputs. Typically, answers to the above two questions are obtained by studying how the system responds
to certain standard inputs. The two most frequently used are the step function input, i.e.,

\[(184) \quad x(t) = \begin{cases} 0, & t < 0 \\ \infty, & t \geq 0 \end{cases},\]

and a velocity step or ramp input

\[(185) \quad x(t) = \begin{cases} pt, & t \geq 0 \\ 0, & t < 0 \end{cases},\]

To answer the question of how fast the system responds requires a rather detailed analysis of the equations. The task of determining the steady state error, however, is often quite simple if one makes use of what it called the final value theorem for Laplace transforms. This theorem states that

\[(186) \quad \lim_{t \to \infty} f(t) = \lim_{s \to 0} sF(s).\]

In many cases, the error \( \epsilon \) will be \( x - y \). Thus the steady state error will be

\[(187) \quad \lim_{s \to 0} \left[1 - H(s)\right] X(s)\]

This is very easy to compute without actually solving the differential equation.

Feedback is a very important characteristic of most regulating systems. It has been noted above that a system will often operate by trying to reduce to zero the difference between the actual and the desired value of some variable. This is typical of feedback systems since it is the error that actuates the system. In a feedback system the actual output is compared with the desired output, and this difference activates the system. A convenient way to represent feedback
control systems is through the use of block diagrams. The most general single feedback loop control system can be represented as shown in Fig. 13. In a block diagram note that the Laplace transforms

\[ X(s) \quad + \quad \varepsilon(s) \quad \rightarrow \quad G_1(s) \quad \rightarrow \quad Y(s) \]
\[ W(s) \quad \rightarrow \quad G_2(s) \]

Fig. 13 — A general single feedback loop control system

of the variables are used. A branch represents a variable. The following figure

\[ Y(s) \quad \rightarrow \quad Y(s) \]
\[ Y(s) \]

illustrates a pickoff point. The same variable coming into the dot is transmitted along each of the branches emanating out from the dot. To represent an adder or differencer, we use the following type of symbol:

\[ X \quad + \quad \varepsilon = X - W \]

Passing through a block, a variable is transformed. The appropriate transfer function is represented in the block.
Thus, for

\[ \epsilon \quad \rightarrow \quad G_1(s) \quad \rightarrow \quad Y(s) \]

it must be true that \( Y(s) = G_1(s) \epsilon(s) \). The equations for the system represented by Fig. 13 can thus be written down as

\[ (188) \quad \epsilon = X - W ; \quad W = G_2 Y ; \quad Y = G_1 \epsilon , \]

or

\[ (189) \quad Y = G_1 X - G_1 W = G_1 X - G_1 G_2 Y , \]

or

\[ (190) \quad Y = \frac{G_1}{1 + G_1 G_2} X = H X , \]

where

\[ (191) \quad H(s) = \frac{G_1(s)}{1 + G_1(s) G_2(s)} . \]

\( H(s) \) is the open loop transfer function for the system. In terms of \( H \), a block diagram for the system becomes that shown in Fig. 14.

\[ X(s) \quad \longrightarrow \quad H(s) \quad \longrightarrow \quad Y(s) \]

**Fig. 14 — Open loop transfer function for the system**

Block diagrams provide a convenient way to obtain the overall transfer function, or, equivalently, the differential equation for the system without using the clumsy sorts of substitutions incorporated in the
previous example. We will not consider here the rules which are available for reducing block diagrams. The block diagram for the inventory-production scheduling example above is that shown in Fig. 15. The feedback character of the system is clearly evident. The

![Block diagram for factory warehouse example](image)

Fig. 15 — Block diagram for factory warehouse example

diagram incorporates the fact that the Laplace transform of \( \int f dt \) is \( F/s \).

Another diagram, the signal flow type, is often useful. It is essentially the dual of a block diagram. The signal flow diagram is a graph and consists only of nodes connected by directed branches. The nodes refer to variables and the branches represent transfer functions. The variable represented by a node is the sum of everything corresponding to branches directed into the node. This variable is transmitted, properly transformed, along branches directed out of the node. The signal flow graph for Fig. 13 is shown in Fig. 16.
Fig. 16 — Signal flow graph for Fig. 13

Similarly, the signal flow graph for Fig. 15 is that shown in Fig. 17.

Fig. 17 — Signal flow graph for Fig. 15

Rules are also available for reducing signal flow graphs to obtain the open loop transfer function for the system, but they will not be considered here.

This section has reviewed the techniques used by electrical engineers in analyzing feedback control systems for deterministic inputs. These techniques also provide the basis for designing such systems. Although the above techniques appear not to have been used extensively in the study of inventory systems it seems clear that they should be quite useful, especially for studying complex multi-echelon
systems. Many of the techniques could be carried over directly with little or no modification. Others, for reasons such as those indicated above, might need to be modified or extended somewhat.

In almost all of the preceding discussion, the assumption was that the variables could be treated as continuous. This approach is most valid in situations where a considerable amount of aggregation occurs, so that the effects of discreteness are smoothed out. For example, in a multi-echelon production distribution system one might aggregate over all outlets at the retail stage and over all stocking points at the wholesale stage.

It should be remembered that only linear systems were studied above. No general results have been obtained for nonlinear systems. For many applications, it would seem reasonable to linearize the actual behavior of an inventory system. In certain cases where large fluctuations are unavoidable, the nonlinear behavior may need to be included.

**DYNAMIC BEHAVIOR - STOCHASTIC INPUTS**

The previous section has dealt with the stability and response of dynamic systems under the assumption that the inputs were deterministic. In this section we will consider stochastic inputs. Clearly, the stability of the system has nothing to do with whether the inputs are deterministic or stochastic. The stable system as defined in the previous section will be stable for stochastic inputs as well. However, for stochastic inputs, the method of handling the response of the system must be extended somewhat.
First, let us examine the techniques for describing the statistical properties of the input time series. With a linear system, if the input can be written \( x(t) = \sum_{v=1}^{r} \xi_v x_v(t) \), then the output can be written
\[ y(t) = \sum_{v=1}^{r} \xi_v y_v(t) \]
where \( y_v(t) \) is the response of the system if \( x_v(t) \) was the only input. This is called the superposition principle. It will be imagined that the input signal can be represented as the sum of three components
\[
(192) \quad x(t) = x_T(t) + \alpha(t) + \xi(t)
\]
In Eq. 192, \( x_T(t) \) is assumed to be deterministic and represent some long term trend. For example, \( x_T(t) = t \) or \( x_T(t) = t^2 \). Then \( \alpha(t) \) is assumed to represent some seasonal or longer term fluctuation about the trend; \( \alpha(t) \) may be deterministic or may be a random variable. For example, if \( \alpha(t) \) were deterministic, it might have the form \( \alpha(t) = A \sin \omega t \). It will be assumed that the average value over time of \( \alpha(t) = 0 \), i.e., the trend is in \( x_T(t) \). Finally, we will assume that \( \xi(t) \) is a random variable which represents "noise" in the input, i.e., random disturbances which are not part of the "true signal."

Consider then the method for describing \( \alpha(t) \) (when \( \alpha(t) \) is a stochastic variable) and \( \xi(t) \). To describe a random time series \( \xi(t) \) one needs a whole collection of probability densities \( f(\xi; t) \), \( f(\xi_2 | \xi_1; t_1, t_2) \), etc., where \( f(\xi; t) \) \( d\xi \) is the probability that the random variable lies between \( \xi \) and \( \xi + d\xi \) at time \( t \), \( f(\xi_2 | \xi_1; t_1, t_2) \) \( d\xi_2 \) is the probability that the random variable lies between \( \xi_2 \) and \( \xi_2 + d\xi_2 \) at time \( t_2 > t_1 \) given that it had the value \( \xi_1 \) at time \( t_1 \). The usual assumption in electrical engineering is that the stochastic processes are stationary,
i.e., the nature of the process generating the random variable does not change with time, so that \( f(\xi_1; \tau) \) is independent of \( \tau \), and \( f(\xi_2 | \xi_1; \tau_1, \tau_2) \) depends only on \( \tau_2 - \tau_1 \). This assumption will be made in the following.

The two most useful functions for characterizing a random time series are the autocorrelation function and the power spectral density. The autocorrelation function for \( \xi(t) \) is defined to be

\[
R_{\xi}(\tau) = \lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{T} \xi(t) \xi(t+\tau) dt
\]

\( R_{\xi}(\tau) \) is the average over all time of the product of values of \( \xi \) separated by a time \( \tau \). When these values are uncorrelated, they will average to zero. If it follows that \( \xi(t+\tau) \) will tend to be high if \( \xi(t) \) is high, then \( R(\tau) \) will be positive, etc.

To define the power spectral density, let

\[
\Omega_{\xi}(i\omega) = \int_{-T}^{T} \xi(t) e^{-i\omega t} dt.
\]

Then the power spectral density is defined to be

\[
P(\omega) = \lim_{T \to \infty} \frac{1}{2T} | \Omega_{\xi}(i\omega) |^2.
\]

Physically, \( P(\omega) \) gives information about how fast \( \xi(t) \) can change, since it says something about the frequencies that are present in \( \xi(t) \). If \( \xi(t) \) contains only low frequencies (i.e., \( P(\omega) = 0 \) at high frequencies), then \( \xi(t) \) cannot usually change too rapidly. On the other hand, if \( \xi(t) \) contains high frequencies then it can change much more rapidly on the average.

The functions \( R_{\xi}(\tau) \) and \( P(\omega) \) are not independent. In fact, \( P(\omega) \) is the Fourier transform of \( R_{\xi}(\tau) \), if the Fourier transform of a
\( f(\tau) \) is defined to be

\[
F(\omega) = \int_{-\infty}^{\infty} f(\tau) e^{-i\omega \tau} d\tau.
\]

It might be noted that deterministic functions without trends also possess autocorrelation functions and power spectral densities. For example, if \( \alpha(t) = a \sin (\omega_0 t + \phi) \), then \( \mathcal{R}_\alpha(\tau) = \left( \frac{\omega_0^2}{2} \right) \cos \omega_0 \tau \), and \( \mathcal{P}_\alpha(\omega) = \pi a^2 \delta(\omega - \omega_0) \), where \( \delta(\omega - \omega_0) \) is the Dirac delta function. Consequently, the functions \( \alpha(t) \) and \( \epsilon(t) \) defined above will always have autocorrelation functions and power spectral densities. However, a trend such as \( x(t) = 0 \), \( t < 0 \), \( x(t) = t \), \( t > 0 \) has neither autocorrelation function nor spectral density since the integrals do not converge. This is the reason that the trend was separated out above.

The cross correlation function between \( \alpha(t) \) and \( \epsilon(t) \) is defined as

\[
\mathcal{R}_{\alpha \epsilon}(\tau) = \lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{T} \alpha(t+\tau) \epsilon(t) dt
\]

(196)

Often it is assumed that \( \alpha, \epsilon \) are uncorrelated so that \( \mathcal{R}_{\alpha \epsilon}(\tau) = 0 \). This will be assumed to be the case in the following.

An important question concerning the operation of any system is what happens to noise or to any random time series as it passes through the system. This is important, because it is a well known fact that purely random fluctuations at the input point can lead to oscillations in the output. In terms of an inventory system this would mean that because of the random character of the demands, even though the mean demand was not changing with time, production might actually exhibit an undesirable cyclical behavior precisely because of the way the noise interacted with the transfer function of the system.
Let the response of the system to \( x(t), \alpha(t), \epsilon(t) \) be denoted by \( y_x(t), y_\alpha(t), y_\epsilon(t) \), so that the output \( y = y_x + y_\alpha + y_\epsilon \). Now we will show how to compute the autocorrelation functions and spectral densities for \( y_\alpha, y_\epsilon \text{ and } y_\alpha + y_\epsilon \). As before, let \( h(t) \) denote the inverse Laplace transform of the transfer function \( \mathcal{H}(s) \). Consider the input \( \epsilon(t) \). Then

\[
(197) \quad y_\epsilon(t) = \int_{-\infty}^{t} h(t-\zeta) \epsilon(\zeta) d\zeta = \int_{-\infty}^{t} h(t-\zeta) \epsilon(\zeta) d\zeta = \int_{-\infty}^{t} h(\zeta) \epsilon(t-\zeta) d\zeta
\]

when \( \epsilon = 0, t < 0, h(t) = 0, t < 0 \). But then, the autocorrelation function \( R_{y_\epsilon}(\tau) \) for \( y_\epsilon(t) \) is

\[
R_{y_\epsilon}(\tau) = \lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{T} y_\epsilon(t+\tau) y_\epsilon(t) \, dt
\]

\[
= \lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{T} \int_{-\infty}^{\infty} h(\zeta) \epsilon(t+\tau-\zeta) \epsilon(t-\eta) \, d\zeta \, d\eta \, dt
\]

\[
= \lim_{T \to \infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(\zeta) h(\eta) \left[ \frac{1}{2T} \int_{-T}^{T} \epsilon(t+\tau-\zeta) \epsilon(t-\eta) dt \right] d\zeta d\eta
\]

\[
(198)
\]

which is the desired result. Equation 198 shows how to compute \( R_{y_\epsilon} \) from \( R_\epsilon \). The equation relating \( R_\alpha \text{ and } R_{y_\epsilon} \) has, of course, precisely the same form.

Consider next the problem of relating \( P_\epsilon(\omega) \) and \( P_{y_\epsilon}(\omega) \). Since \( P_{y_\epsilon}(\omega) \) is the Fourier transform of \( R_{y_\epsilon}(\tau) \), it follows that
\[ P_y(e^{j\omega}) = \int_{-\infty}^{\infty} R_y(\tau) e^{-j\omega \tau} d\tau = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(\zeta) h(\eta) R_\epsilon(\eta+\zeta) e^{-j\omega \tau} d\zeta d\eta d\tau \]

\[ = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-j\omega (\eta+\zeta)} e^{j\omega \eta} e^{-j\omega \zeta} h(\zeta) h(\eta) R_\epsilon(\eta+\zeta) d\zeta d\eta d\tau \]

\[ = \left[ \int_{-\infty}^{\infty} e^{j\omega \eta} h(\eta) d\eta \right] \left[ \int_{-\infty}^{\infty} e^{-j\omega \zeta} h(\zeta) d\zeta \right] \left[ \int_{-\infty}^{\infty} e^{-j\omega (\eta+\zeta)} R_\epsilon(\eta+\zeta) d\tau \right] \]

\[ \tag{199} = H(-j\omega) H(j\omega) P_\epsilon(e^{j\omega}) = \left| H(j\omega) \right|^2 P_\epsilon(e^{j\omega}) , \]

which is the desired result. There is, of course, a similar equation for \( P_y(e^{j\omega}) \).

If \( y_1 = y_\alpha + y_\epsilon \), then when \( \alpha \) and \( \epsilon \) are uncorrelated

\[ \tag{200} R_{y_1}(\tau) = R_{y_\alpha}(\tau) + R_{y_\epsilon}(\tau) , \]

and

\[ \tag{201} P_{y_1}(e^{j\omega}) = P_{y_\alpha}(e^{j\omega}) + P_{y_\epsilon}(e^{j\omega}) . \]

Given the above information, the basic problem is to design an optimal system. In electrical engineering, optimum is almost always interpreted to mean a system which minimizes the mean square error.

It will be assumed that the control system has no difficulty in following properly the trend \( x_\alpha(t) \), and hence the design problem centers around the input \( x_1(t) = \alpha(t) + \epsilon(t) \). The error at any point in time might be \( E = x_1(t) - y_1(t) \). More often, it will be \( E(t) = f(t) - y_1(t) \), where \( f(t) \) is obtained from \( x_1(t) \) by a combination of operations such as integration or differentiation of \( x_1 \). For example, in the context of inventory problems, \( x_1 \) may be passed through an nth order exponential
delay to yield \( f(t) \) so that \( f(t) \) may be something like the desired inventory level whereas \( y_1(t) \) is the actual inventory level. The problem then is to determine the transfer function for the system which minimizes

\[
I = \lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{T} E^2(t) \, dt.
\]

(202)

To see how to determine the optimal transfer function, note that

\[
y_1(t) = \int_{0}^{t} h(t-\zeta) x_1(\zeta) \, d\zeta,
\]

so

\[
I = \lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{T} \left[ f(t) - \int_{0}^{t} h(t-\zeta) x_1(\zeta) \, d\zeta \right]^2 \, dt,
\]

\[
= \lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{T} \left[ f^2(t) - 2f(t) \int_{0}^{t} h(t-\zeta) x_1(\zeta) \, d\zeta + \left( \int_{0}^{t} h(t-\zeta) x_1(\zeta) \, d\zeta \right)^2 \right] \, dt
\]

\[
= R_f(0) - 2 \int_{0}^{t} h(\zeta) R_{fx_1}(\zeta) \, d\zeta + \lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{T} \int_{0}^{t} \int_{0}^{t} h(t-\zeta) h(t-\rho) x_1(\zeta) x_1(\rho) \, d\zeta \, d\rho \, dt
\]

\[
= R_f(0) - 2 \int_{0}^{t} h(\zeta) R_{fx_1}(\zeta) \, d\zeta + \int_{0}^{t} \int_{0}^{t} h(\zeta) h(\rho) R_{x_1}(\zeta-\rho) \, d\zeta \, d\rho
\]

(203)

\[
= R_f(0) - \int_{0}^{t} \left[ 2 R_{fx_1}(\zeta) - \int_{0}^{t} h(\rho) R_{x_1}(\zeta-\rho) \, d\rho \right] h(\zeta) \, d\zeta
\]

Now we must determine the function \( h(\zeta) \) which minimizes Eq. 203. This is a problem in the calculus of variations. If \( h^*(\zeta) \) is the optimal function, and variations of the form \( h(\zeta) = h^*(\zeta) + \sigma \delta(\zeta) \), then \( I \) becomes a function of the parameter \( \sigma \), and a necessary condition which \( h^*(\zeta) \) must satisfy is \( I'(0) = \left( \frac{dI}{d\sigma} \right)_0 = 0 \). Now
\[ I'(0) = -2 \left[ \int_0^t R_{x_1}^*(\zeta) - \int_0^t h^*(\rho) R_{x_1} (\zeta-\rho) d\rho \right] \delta(\zeta) d\zeta \]

This must hold for all admissible variations \( \delta \), and hence by the fundamental lemma of the calculus of variations, \( h^*(\zeta) \) must satisfy the equation

\[ \int_0^t h(\rho) R_{x_1} (\zeta-\rho) d\rho = R_{x_1}^*(\zeta). \]

Equation 205 is an integral equation whose solution yields \( h^*(\zeta) \).

It is often referred to as a Wiener-Hopf integral equation. It is possible to solve this equation, but usually only in relatively simple cases. Electrical engineering frequently makes use of the theory just developed to determine the optimal linear predictor for a noisy signal.

The preceding is a brief survey of the approach which electrical engineering uses with design systems when the inputs are stationary random variables. Although the methods are clearly relevant to inventory problems, very little has been done to apply them in this area. Certainly, the technique could not be used directly with a single stocking point which orders not continuously but in discrete amounts. It might be used, however, for an aggregated production distribution system, where the output was the production rate. It would be necessary to change the criterion, however, from minimizing the mean square error to minimizing an appropriate cost function. In a very limited number of cases, the above techniques have been used (in discrete form where the integral equation becomes a finite set of linear equations) to determine the best linear predictor for forecasting demands. In this case the minimization of the mean square error can be used. The problem is to find the optimal weights \( \beta_i \).
such that if

\[ D'_{j+k} = \sum_{i=0}^{N} \beta_i D_{j-i} \]

is the predicted demand for period \( j + k \), then \( \sum (D_{j+k} - D'_{j+k})^2 \) is minimized. In order to do this, of course, it is necessary to know the autocorrelation function for demand. This is often hard to determine in practice. 

* A more detailed treatment of the material in the last two sections can be found in Ref. 39.
VARIOUS USES FOR SIMULATION

Simulation has already been mentioned as another useful tool for studying inventory systems. Simulation can substitute for, complement, or supplement analytical or heuristic-intuitive analysis. There are a variety of ways in which simulation has been or can be used in studying inventory systems. These may be conveniently classified according to how simulation is used: (1) as a complement to analytical analysis; (2) as a complement to heuristic-intuitive analysis; or (3) as an alternative to the analytical approach.

Consider first the uses as a complement or supplement to analytical analysis. Often simulation can provide a graphic portrayal of the way in which some operating doctrine obtained from analytical models will behave when installed in the real world system. The results are frequently presented by plotting as a function of time the onhand inventory obtained through simulation. The simulation procedure will also be used frequently to compare the new operating rules with the ones previously in use. Here the system is simulated once using the new rules and once using the old ones when precisely the same pattern of demand is employed in both. The comparison can be made by computing the costs incurred under the two operating doctrines over the length of time for which the simulations were carried out. (30)

Another way in which simulation can be used along with analytical methods is to study parameter variations or to make sensitivity
analyses which are difficult to do analytically, or to study the behavior of the operating doctrine obtained analytically under conditions which could not easily be taken into account analytically. For example, in the analytical model it might have been necessary to assume that demands in any non-overlapping time periods were independent, and simulation designed to show what happens as the nature of the autocorrelation function for demand is changed.

The uses of simulation in conjunction with analytical work are usually for relatively simple systems consisting of a single outlet -- perhaps a single warehouse or one department in a retail store. Analytical models do not exist for very complex systems. When more complex systems are simulated, the process is usually used as a substitute for analytical analysis or in conjunction with the heuristic-intuitive approach.

The next category of uses for simulation operates in conjunction with the heuristic-intuitive approach to the development of operating doctrines for inventory systems. The systems studied here will often be rather complex ones. Usually in these cases it is interesting, using simulation, to see how the system might perform under a specified complete set of operating rules for the system (or for some subsection of the system); or it is illuminating to compare two or more different sets of operating rules. The operating doctrines in these cases will have been obtained mainly from a heuristic-intuitive analysis although analytical approaches may have been used for certain small subsegments. In general, it is quite difficult to devise a complete set of rules for a very complex system which will even qualitatively
give the desired type of behavior under a wide variety of conditions. In such situations, no attempt is being made to optimize the system. Indeed, it is usually very difficult even to define what is meant by optimal. One is merely looking for a set of rules which seem efficient, or for the better of two or more sets of rules. (31)

In cases of this sort, simulation often gives the formulators useful insights and a deeper understanding of the system's operation. Then, too, it reveals inadequacies or inconsistencies in the operating rules that might not have come to light otherwise unless the rules were actually implemented in the real world system. On the other hand, it often turns out that the results of the simulation are difficult to evaluate. It is hard to represent the real world in complete detail. As a result, when one devises a simulation, he must make certain simplifications and approximations. In so doing, one must be very careful to see that the resulting system contains enough of reality to adequately test whatever is being examined. It is of course difficult to subject the system in the simulation to all the possible types of stresses which can be encountered in the real world. In addition, a set of criteria may not be available to evaluate systems and to select between alternative operating doctrines. Finally, when one performs a simulation it is often necessary to give parameters numerical values which are hard to estimate in the real world. The outcome of the simulation may or may not be sensitive to the values of these parameters. The problem is that when there are a number of such parameters, it is very difficult to make sensitivity analyses of all of them. All these problems make it
difficult in many cases to derive any clear-cut conclusions from the results of a simulation.

A final category involves simulation as an alternative to the analytical approach. For example, one might use simulation to find the optimal values of one or more parameters in an operating doctrine, in such a way as to minimize some cost expression. A specific case might involve an attempt to determine the optimal values of K and k for a periodic review system. In general, this use of simulation tends to be very inefficient and expensive, and therefore not widespread.

One of the main uses of simulation as an alternative for the analytical approach is in the study of the stability and dynamic response of complex systems. The theoretical foundations for these subjects was covered in pp. 97 - 125. For these purposes simulation is a valuable and relatively efficient tool. The next section will be devoted to studying how simulation can be used to study the stability and dynamic response of systems.

SIMULATION AND THE DYNAMIC BEHAVIOR OF SYSTEMS

We have noted earlier that economic systems, just like electro-mechanical systems, can be unstable. Typically, instability is exhibited by an oscillatory behavior of the system, which is usually undesirable because it prevents the system from operating effectively. Unstable oscillations can be excited in the system by shocks in the inputs or by the nature of the random noise pattern of the inputs. The character of the unstable oscillations, i.e., the amplitude and frequency, is determined to a large extent by the nature of the system itself, especially by the time lags. The stability of any inven-
tory system can be studied with the aid of simulation. It is also possible to use simulation as an aid in studying ways to stabilize an unstable system, or to damp out oscillations.

Another problem concerning the dynamic behavior of a system is the way it responds to a sinusoidal input -- say a seasonal variation in demand. As they pass through the system, both the amplitude and phase of the sinusoidal input can be changed. Let us consider, for example, a production distribution system. A seasonal pattern of demand at the retail level can result in a seasonal pattern of orders at the factory with a different phase and amplitude. In certain situations the system will increase the amplitude of the input for a certain range of frequencies. This is an unfortunate circumstance since it can mean, in the context of the example given above, that the amplitude of oscillations in the ordering rate can be greater at the factory than at the retail level. A change in the phase of the oscillations can be either desirable or undesirable, depending on the circumstances. Simulation can be used to study problems of this sort as well. Furthermore, it can aid in determining how to change the nature of the system response if this is necessary.

A final problem involving the dynamic behavior of systems is the capacity of the system for following a changing input. This is important in inventory systems, since a system which cannot respond quickly to changes in the rate of demand may not be acceptable. There is a close connection between the responsiveness of the system and its stability. Often a system which can respond very rapidly will be unstable. A stable system will frequently be somewhat sluggish
in its response to changes. A compromise is always necessary between the responsiveness of a system and its stability.

The means by which the above problems can be investigated with the aid of simulation will be illustrated in the context of a simple example. Consider a retailer who orders shirts each week to replenish his stock. Delivery takes precisely three weeks from the time the order is placed. He determines his inventory only at the end of the week after any orders arriving that week are on hand. The retailer attempts to maintain an onhand inventory equal to the last five weeks' sales.

In order to employ simulation in studying the situation, one must have a set of equations which describe the operations. These can be classified as decision equations, material balance equations, delay equations, and auxiliary equations which provide definitions of certain variables, etc. For the simple system under consideration there is only a single decision equation, which tells how much the retailer orders at the end of the week. This decision rule can take on many forms. For our purposes here, a rule similar to that often found in practice will be used. Let \( Q_n \) be the quantity ordered at end of week \( n \), \( S_n \) the demand in week \( n \), \( \hat{I}_n \) the desired onhand inventory at the end of week \( n \), and \( I_n \) the actual onhand inventory. Assume then that the retailer determines \( Q_n \) using the rule

\[
Q_n = \begin{cases} 
S_n + \alpha (\hat{I}_n - I_n) & \text{if } (\hat{I}_n - I_n) > 0 \ (0 \leq \alpha \leq 1) \\
0 & \text{otherwise}
\end{cases}
\]
The number $\alpha$ tells how fast the retailer attempts to bring the actual inventory level up to the desired inventory level.

Let $R_n$ be the number of units added to inventory from the receipt of procurements in week $n$. Then the material balance which relates the onhand inventory at the end of week $n$ to the onhand inventory at the end of week $n-1$ is

$$I_n = \begin{cases} I_{n-1} + R_n - S_n & \text{if } > 0 \\ 0 & \text{otherwise} \end{cases}$$ \hfill (208)

It will be assumed that demands occurring when the system is out of stock are lost.

There is only a single delay equation which relates $Q$ and $R$.

It is

$$R_n = Q_{n-3}$$ \hfill (209)

Finally, there is one auxiliary equation which defines $\hat{I}_n$. It is

$$\hat{I}_n = S_n + S_{n-1} + S_{n-2} + S_{n-3} + S_{n-4}$$ \hfill (210)

Using Eqs. 207 through 210, one can compute the behavior through time of the retailer's system for any pattern of sales. Note that this series constitutes a set of nonlinear difference equations. They are nonlinear because of the non-negativity restrictions in Eqs. 207 and 208.

The above simple system has been formulated in discrete terms. If one is simulating a complex production distribution system where there are hundreds of retailers and many distributors, it will usually
be desirable to aggregate over the retailers and similarly over the
distributors. Then one will have a single ordering equation des-
cribing the behavior of all retailers. In the above, the time inter-
val Δt was one week, the period between placing orders. No time
interval of less than one week was of any consequence. When aggrega-
tion is carried out, it is usually convenient to imagine that time
is continuous and the difference equations become differential equa-
tions. The quantity ordered then becomes an ordering rate Q, and
Eq. 207 would become

\[
Q = \begin{cases} 
S + \alpha (\hat{I}_n - I_n) & \text{if } 0 \\
0 & \text{otherwise} 
\end{cases}
\]

where S is now the sales rate. Similarly, Eq. 208 becomes

\[
\frac{dI}{dt} = R - S,
\]

where R is the rate of receipt of units. To solve such a set of
differential equations on a computer, one must again represent them
as a set of difference equations. Now, however, the size of the
time step will appear explicitly and there is no natural time step.
Thus Eq. 212 might be written

\[
I_{n+1} = I_n + \Delta t(R_n - S_n),
\]

where Δt is the size of the time step taken. Equation 213 is only
one possible representation of Eq. 212 as a difference equation.
An infinite number of possible representations exists. It might be
mentioned, as an interesting sidelight, that the size of the time step
and the way in which the differential equations are represented are very important factors in making the simulation. It is well known that difference equations can exhibit instabilities of their own which are not a characteristic of the differential equations that they are representing. Consequently, if one is not careful, the results of the simulation may exhibit oscillations which are not characteristic of the system being simulated, but are instead oscillations introduced by the difference equations which represent the system. If this behavior is to be avoided, the time step will normally have to be considerably smaller than the smallest time delay appearing in the system. This places a serious restriction on the upper limit to the time step. Often it tends to force one to aggregate several small time delays into a single longer one in order to prevent the simulation from requiring an inordinately long time on the computer.

A typical procedure for testing the stability of a system is to determine its response to a step function, which provides a shock to the system that excites the unstable modes of operation. The procedure can be illustrated in terms of the simple example presented above. Suppose that in the past shirts have been selling at the constant (deterministic) rate of fifty per week. Now imagine that sales suddenly jump to a constant rate of seventy-five per week and remain at this value. This is a step function change in the sales rate. Using Eqs. 207 through 210, one can compute the way the system responds to this change through time. The nature of the response will depend critically on the value of $\alpha$ in Eq. 207.
This is illustrated in Fig. 18 where the on hand inventory is plotted as a function of time (the step function change in sales is assumed to occur at time 0). It will be noted that when \( \alpha = 1 \), the system is unstable and never settles down; when \( \alpha = 0.5 \) the system is stable and settles down following an oscillatory path; when \( \alpha = 0.25 \) the system is almost "critically damped" and settles down rather quickly; when \( \alpha = 0.1 \) the system performs sluggishly and slowly climbs to the desired value; and finally when \( \alpha = 0 \), the new desired inventory is never attained, since no attempt is made to correct it. The very simple above example shows that as the retailer attempts to respond more and more rapidly the system changes from a sluggish one to an unstable one. Because of the time lag in delivery of orders
there is a limit to how rapidly the retailer can respond. For the simple model just considered, it is quite easy to work this solution out by hand. However, for more complex systems, it becomes impossibly laborious to carry out the operations by hand and the aid of a computer is needed.

For the simple example just considered, it is not difficult to eliminate the instabilities illustrated in Fig. 18. If instead of basing his decisions on the on-hand inventory level, the retailer uses the quantity on hand plus on order, the oscillations disappear even for \( \alpha = 1 \). Oscillations in the simple system studied were caused by the fact that the decision maker did not take account of what was in the pipeline. In more realistic situations, where there is coupling among the retailer, wholesaler, and the factory, the removal of oscillations is not nearly such a simple matter, because there can be coupled oscillations between the various levels -- retailers, wholesalers, and the factory, just as with coupled pendulums in mechanics, i.e., in mechanical terms there is an exchange of energy among the three levels, which leads to oscillatory behavior.

The same sort of procedure described above can be used to test the response of the system to a sinusoidal input. By changing the frequency one can determine how the phase shift and attenuation vary with the frequency. For linear systems, the attenuation (ratio of output amplitude to input amplitude) is independent of the amplitude. This is approximately true for nonlinear systems -- again, provided that the nonlinear cutoffs are not reached. For nonlinear systems, the output need not be precisely sinusoidal even though the input is.
There is one special problem that one is faced with in studying the sinusoidal response of a system by simulation. About the only feasible way to initiate simulation to start things off is to have the system in steady state, with nothing changing with time (for complex systems, incidentally, it can be quite difficult to determine all the steady state values in the system for some specified inputs). Then, the state being steady, the sinusoidal input is imposed on the system. A transient behavior will occur while the system is adjusting itself to the new input. One must wait until the transient period is over before attempting to evaluate the response of the system. Often it is hard to decide ahead of time how long the transient period, which may be short or long, will last. In addition, it is not always easy to determine by looking at the results whether or not the transient period has ended.

Several approaches can be used to study the way the system responds to a changing input. One important measure is the steady state error, viz., the ultimate difference between the desired and actual inventory. No system can follow an arbitrary input without steady state error, but an increase in the complexity of the system's operating rules can make it follow more and more complex inputs with no long-run error. Consider as an example the simple system which we have been using. When the system is stable, the actual inventory will be equal to the desired inventory at $t = \infty$ except when $\alpha = 0$. When $\alpha = 0$, the actual inventory will always be less than the desired inventory; then there will be a steady state error. That there should be such an error is obvious since no attempt is made to correct the inventory level.
Now imagine that the sales rate has been constant at fifty shirts per week in the past, and, at \( t = 0 \), the sales rate begins to increase at a constant five units per week. After an initial transition phase, the desired inventory will increase at the rate of 25 units per week. The actual inventory for \( \alpha > 0 \) and \( \alpha \) small enough for the system to be stable will also ultimately increase at the rate of 25 units per week. However, when \( t = \infty \) there will be a fixed differential between the desired and actual inventories. This is shown in Fig. 19. The reason for this steady state error, of course, is simply that the retailer does not have built into his decision rule any scheme to forecast future sales. With a forecasting system, it would be possible to reduce the steady state error in the above case (a ramp input or velocity step as it is called) to zero. It might be pointed out, however, that the introduction of

![Fig. 19—Response to a ramp input](image-url)
forecasting rules will usually increase the tendency of the system
to be unstable.

The inputs to the system described thus far are what might be
called deterministic. In the real world, noise (i.e., random
fluctuations) is also often present in the input. Simulation is
useful in studying the way in which various sorts of noise patterns
can influence the system. The noise pattern can be conveniently
characterized by its autocorrelation function or its power spectral
density (the Fourier transform of the autocorrelation function).
Both of these measures tell how rapidly the input can change from a
given value to a different value. If the frequency spectrum does
not contain high frequencies, then it is difficult for the input to
change rapidly. On the other hand, if high frequencies are present
the input can change its value quickly. It is possible on a computer
to generate random numbers with different kinds of autocorrelation
functions; hence one can study the effects of various sorts of noise
inputs on the system.

Like an electro-mechanical system, an inventory system can have
a natural frequency. If the input has a frequency close to the
natural one, any amplifications which take place will be considerably
exaggerated compared to other frequencies. The natural frequency of
the system can also be excited by the noise pattern of the input, and
then one can find oscillations in the system whose frequency is the
natural frequency even though none of the inputs has a frequency
even close to the natural. Simulation can be used to point up the
occurrence of such behavior.
The above discussion has pointed out some ways in which simulation can be used as an aid in studying the dynamic behavior of systems. Often, of course, the purpose of such analyses is to discover any undesirable behavior and attempt to eliminate it. It is not always easy to eliminate such behavior without making radical changes in the way the system operated, since it is necessary to change various time lags, operating doctrines, or both.

Simulation is a relatively recent procedure in studies of the dynamic response of systems. Perhaps the most widely publicized effort along these lines has been that done under the name of Industrial Dynamics. (32)(33)
Chapter 4

HEURISTIC-INTUITIVE APPROACH - ART OR SCIENCE?

Basically, all the ideas for analytical models or simulations arise as a result of heuristic or intuitive considerations. It is possible, however, to design and operate inventory systems using only the heuristic-intuitive approach based on experience. This approach may be termed inventory management as an art rather than as a science. In the real world, including industry, commerce, and the military, the great majority of inventory systems now in operation or being installed have been developed and are being run on the basis of intuition and experience alone. In certain circumstances, this approach seems to yield results at least as good as could be obtained using a more scientific approach. This merely serves to point out that in certain situations, inventory management as a science has not yet caught up with inventory management as an art.

In many situations, however, it is apparent that something more than the heuristic-intuitive approach is needed. This is especially true for very large complex systems such as those operated by the military. At first, one might think that the heuristic-intuitive approach would be best for the military, since no analytic models are available to represent in any detail the complex systems involved. The difficulty is that no single person can be familiar in complete detail with the operation of the entire system. Hence the procedure has been to let different individuals or groups determine the operating rules for various segments of the system. Inasmuch as the members of one group are only vaguely familiar with the way the rest of the
system is to operate, the resulting sets of operating rules may be at odds with one another, or completely inconsistent, or notably inefficient when used together. The installation of new complex inventory systems (or any other systems, for that matter) has never been an easy task; often years go by before anything close to satisfactory operation is attained. The problems have been brought clearly into focus recently when the military attempted to install several complex inventory systems (to be used with new weapon systems) and hoped to have them operating efficiently in a very short time. None of the attempts was a complete success; at least one bordered on being a total failure initially, and was only partially operating several years after installation. To have any reasonable expectation that a new complex system can be put into operation in a short amount of time, one must make a detailed study of the behavior of the individual operating rules and the way they integrate together, of the flows of materials and information in the system, of the jobs which humans are to perform, and of the set of incentives active in the system. Such an analysis will require a scientific approach to the problem and will generally require a blending of intuitive, analytical, and simulation techniques.
Chapter 5

IMPORTANT UNSOLVED THEORETICAL PROBLEMS

MULTI-OUTLET AND MULTI-ECHELON INSTALLATIONS

Almost all the theoretical work on inventory systems has been concerned with stockage at a single installation. In reality, however, a large percentage of the systems have more than a single outlet, and many are multi-echelon. Military logistics systems, for example, provide a good illustration of multi-echelon, multi-outlet systems. There are basically two classes of problems concerned with multi-echelon, multi-outlet systems, neither of which has received much theoretical analysis. The first class imagines that the spatial structure of the system (i.e., the number of stocking points and their echelon arrangement) is given, and seeks to determine the optimal way to operate the system not only in steady state, but with noisy, time varying inputs as well. The second and even more interesting class of problems attempts to answer the questions: "Given the job to be done, how many outlets, how many stocking points, and how many echelons should there be, and what is the optimal policy for operating the resulting system?" The simplest example of this type is the warehouse location problem.

In industry, one can discover a wide variety of inventory systems. Some are single-echelon, some are multi-echelon; some use only a large central warehouse while others use a number of small branch warehouses, etc. In addition, it is possible to find at a single point in time one company shifting from a multi-echelon to a single-echelon system while another is shifting from single to
multi-echelon, or from a central warehouse to branch warehouses, and vice-versa. All firms claim that the changes will reduce costs. At this time it seems that the factors favoring one type of system over another are not precisely known.

It is clear that the second class of problems will probably be quite difficult to solve. Even the first class has resisted the analytical approach because of the various possible policies for shipping from one echelon to another and for redistribution between echelons -- both on a routine or expedited basis. Of course, all sorts of other complications are possible, such as the existence of dependence between the demands at the various outlets. Simulation can be used as an aid in studying multi-outlet, multi-echelon problems. Usually, however, it is limited to studying only two or three possible operating doctrines. Since for a complex multi-outlet, multi-echelon system the number of feasible operating doctrines is huge, it becomes a cumbersome task to use simulation to explore in great detail the various possibilities. Nonetheless simulation can sometimes yield useful insights on how to improve, although not optimize, some system.

PRODUCT INTERACTIONS

Most inventory models study each item in isolation and neglect any interactions which may exist between the given item and the other items carried by the system. In some models, constraints are imposed which allegedly take into consideration the interaction between products. Later we will say more concerning the inadequacies of the various approaches to the constraint problem. Aside from competing for budget allocation, floor space, etc., the various
products interact in a number of ways. The most obvious of these is that the various products are often substitutes for or complementary to one another. These relationships have a substantial effect on demand for a particular inventory item, yet they are not considered in the demand distributions assumed by most authors. Indeed, to consider these interactions explicitly would be impossible from a practical standpoint. Other interactions occur on the procurement side where cost savings can sometimes be accomplished by procuring several items from the same source. These cost savings can result from quantity discounts related to the total volume of purchases or from shipping economies.

In the event that there is a manufacturing process involved, the existence of many products can complicate the problem tremendously. The order in which the items are processed can affect the amount of setup costs incurred for the total number of items. Also, the joint setup times must be taken into consideration in determining the optimal length of run for each product. Sequencing problems have been solved only for extremely simple problems. Research in this area is still in its infancy. The problems are usually of a combinatorial nature and have not yielded to the standard operations research techniques.

**MULTI-STAGE PRODUCTION**

The problem of controlling the inventories of a firm which manufactures a product or products that must undergo a sequence of several production operations with the possibility of carrying inventories of semifinished goods between operations has not received
much attention. In fact we do not know of a single inventory model that can be usefully applied to a multi-stage production operation to set inventory levels at each of the semifinished goods stocking points. Oddly enough, there is little comment in the literature pointing out the inadequacies of inventory models to cope with multi-stage problems, in spite of the fact that these problems are frequently encountered in industry. The analysis has not been completed even for a firm with one product. Typically, inventory models are applied to some particular stage of production with no attempt to analyze the effects of the application on the levels of inventories at other stages. The most typical application is that of the simple lot-size model to minimize the sum of setup costs and inventory carrying charges in the stage that has the largest setup cost.

INTERACTIONS BETWEEN SYSTEM BEHAVIOR AND DEMAND

The usual type of inventory model assumes that the stochastic processes generating demands on the system are independent of the behavior of the system. In the real world, this is not true either in business or the military. In a military supply system for example, if an item reaches short supply, supply officers will tend to stockpile it even though they have no need for it. Similarly in industry, attempts are made to stockpile an item if it is expected to be in short supply in the future.

In industry there is a variety of interactions between the producers and consumers. The price of the item obviously has an important effect on the demand for the item. Inventory models usually attempt to minimize expected costs. Only when the price is fixed is
cost minimization equivalent to profit maximization. Existing research has done little to include something like a demand curve in the theoretical analyses. Similarly, the magnitude, timing, and nature of advertising can influence the level of demand and the way in which it changes over time. These factors are never included in the inventory models either. All this seems to point out that the inventory problem is only a part of the overall problem of operating a firm, and sometimes it is difficult and dangerous to divorce its study completely from the remainder of the firm's operations.

**REDISTRIBUTION ASPECTS**

When more than one outlet is involved, one must consider the possibilities of redistributing items among the outlets. Existence of these possibilities can exert a strong influence on the system inventory policy. Models have sometimes been constructed which ignore these aspects in either of two ways. Some involve the setting of inventory levels for the system as though all items in the system are at all times available at all outlets, i.e., the implicit assumption being that redistribution can be effected freely and instantaneously. Others set inventory levels on an individual outlet basis without allowing for redistribution; i.e., the implicit assumption is that redistribution costs are prohibitively high.

As mentioned above, inventory models have been constructed that take redistribution aspects into consideration.\(^{(27)(35)}\) However, these are not free from possible objections. The models involved the setting of redistribution trigger levels at a constant level, even though it can be argued that the optimal level is in fact a
function of time. For example, if an outlet's stock of an item reaches its redistribution level very early in a period, redistribution is more justifiable than it would be if the stock level reached the same level later. Unfortunately, the solution to the problem of setting redistribution levels as a function of time is not easily accomplished.

STOCHASTIC LEADTIMES

Many inventory models, assuming that demand rate is the only variable subject to random variations, do not give explicit consideration to variations in leadtime. In fact, variations in leadtime can in practice be the cause of more variation in leadtime demand than variations in demand. Assuming that the demand distribution and the distribution of leadtime are independent random variables with means $\lambda t$, $\mu_t$ and variances $\sigma^2_x$, $\sigma^2_t$ respectively, one can demonstrate that the variance of leadtime demand is:

$$\lambda^2 \sigma^2_t + \mu_t \sigma^2_x .$$

This formula is not particularly useful, however, unless it is known that the distribution of leadtime demand is approximately normal.

Other problems concerning the proper representation of variable procurement leadtimes exist, however. Except for queueing models, only a few articles have appeared on the subject of variable leadtimes in inventory control. The usual assumptions made are that leadtimes are independent of the process generating the orders and that leadtimes are independent of each other so that orders can cross and need not be received in the sequence in which
they were placed. In fact, neither of these assumptions is valid. If demands are exceptionally heavy, the result is likely to be that queues will form at the source, resulting in longer leadtimes. Concerning the assumption that orders can cross, it is in fact much more likely that orders will be received in the sequence placed. If the orders do cross, the expected cost formulas misrepresent the costs incurred. Thus, only when the probability that orders can cross is extremely low (leadtimes being considered independent) can the expected cost formulas based on the independence assumption give reasonably accurate results. For an accurate description of the nature of procurement leadtimes, it is necessary to have a combined model of the inventory system and the supplier. The leadtime distribution depends on whether there is only one supplier or whether there are several suppliers, on the nature of the delays at the supplier's facility, and on the contractual arrangements, i.e., penalty clauses, etc. with the supplier. To construct such a model is complicated, both from a theoretical and practical standpoint.

Another type of leadtime problem that can be important in practice but has not been included in theoretical formulation of inventory problems is the fact that an inventory system can be supplied by more than one mode of delivery. Routine and priority modes of resupply can be in operation simultaneously. The inventory literature to date assumes the existence of only one mode of delivery, whereas it is the general rule in practice that more than one mode is utilized. Except for cases in which the alternative mode is instantaneous, no adequate models have been constructed to cope with the existence of more than one means of transportation.
Even if only one mode is used, there can be alternatives from which that one will be selected. The cost of shorter leadtime may be reflected in the unit cost of the product. The trade-off between leadtime and unit cost has been examined for a simple lot-size reorder-point model in a recent report.(38) For the case of low demand items, a model was derived at RAND in 1954 that determines stockage objective and a leadtime which minimizes the expected costs of stocking, running out of stock, and transportation costs.(12)

Finally, it might be pointed out that other complications may arise in practice which make it difficult to treat leadtimes realistically. These include the possibility of negotiating leadtimes for individual orders and for expediting delivery if it appears that the order will be needed sooner than expected. In some industries, the practice of having one company supply the needs of another in case of emergency is also employed.

PERIODIC REVIEW VERSUS TRANSACTIONS RECORDING

Periodic and continuous review models have been discussed above. Both types of systems are used in practice, but no real decisions have been reached concerning which of the two types of system is preferable in the various applications. A rational choice of model necessarily involves comparison of costs (including review costs of the optimized discrete review model); the expense of having the computer space constantly available for recording demands and making decisions when demands occur; and the cost of the optimized continuous review model. Up to this point, few calculations have been made to compare the costs of these two models. For extremely low demand

---

*(1957 ed., Appendix 4)*
items, one would expect lower costs to result from a continuous review model (which only considers an item when transactions actually occur) than would result from a periodic review model. On the other hand, for items with high demand, the data processing costs might be reduced by utilizing a periodic rather than a continuous review model.

REPAIR MODELS

Much of the operations research literature on repair problems consists of queueing models. The interested reader can inform himself about these models in P. M. Morse’s *Queues, Inventories, and Maintenance*. Most of the models apply to problems of repair and preventive maintenance for machines. Arrivals are generated by machine breakdowns and servicing is carried out by one or more repairmen at the machine location. The type of repair problem encountered in most military inventory problems for repairable items is more complicated in several respects. Demands for the item are not all generated by failures; i.e., the end items can be consumed. Not all the items are repairable, although the usual queueing model does not allow for nonrepairable items. Furthermore, there are leadtimes involved as well as repair times, and it is not clear that the service distribution assumptions appropriately describe the real time patterns. Finally, repair is often done at a centralized facility, and it is difficult to build into the queueing models an allocation rule for optimal return of the units to a number of bases.
AGGREGATE VERSUS INDIVIDUAL ITEM CONTROL

We have presented in earlier sections models that were based on analysis of control for individual items and models for determining the level of aggregate production, inventory, and work force. Adequate criteria have still not been developed to determine whether the aggregate or the individual item approach is preferable. The decision is not based on analytical principles. Not a great deal has been accomplished in the reconciliation of results obtained by the two different approaches. The linear programming approach to determining aggregate production and inventory levels does not take into consideration setup costs or probabilistic variations. The linear decision rule model requires that quadratic cost functions provide an adequate approximation of actual cost behavior.

BUDGET CONSTRAINTS

None of the inventory models discussed in operations research literature appears to provide an adequate treatment of budgetary constraints. In the first place, budgets are usually set in terms of actual dollars to be spent, whereas the costs included in inventory control models are often not the out-of-pocket cost type. For example, inventory carrying charges include elements of opportunity cost such as the rate of return on alternative uses of capital, and the costs of placing orders are not typically charged against the procurement budget.

Another type of problem is that budgets are set in terms of a specific number of dollars. Since demand is a random variable, the number of dollars required to keep the system in operation is also
a random variable. Thus, for a given inventory control system, problems arise when the amount budgeted is not sufficient to cover replacement of the units demanded in the period. If this situation persisted, it could be serious enough to cause the system to run down hill and destroy the steady state nature of the model. Another practical difficulty is that when budgeted funds are not completely spent, the budget for the following year is likely to be cut.

Serious budget problems can also arise during the transition from the existing system to a new system if additional funds are needed to effect the change. It is important that special funds be allocated to accomplish the transition; otherwise the new system cannot be put into effective operation.

**LOST SALES MODELS**

Most of the models developed in the inventory literature assume that demands occurring when the system is out of stock are backordered. The two simple models for the lost sales case presented above were valid only under the assumption that the system was out of stock no more than a very small fraction of the time. The only lost sales case that has been solved exactly is that in which only a single order is outstanding and demand is Poisson distributed. It is sometimes possible to specify rigorously in the lost sales case that no more than a single order will ever be outstanding. Such a condition was not possible in the backorders case. The case where more than one order is outstanding has not been solved. The lost sales case represents a fundamentally different stochastic process from the backorder case.
Chapter 6

DIFFICULTIES OF PRACTICAL APPLICATION

THE PROBLEMS

Even if all the theoretical problems discussed above were solved, many practical difficulties would have to be overcome in any given case before a successful implementation could be made. We have not attempted in this report to describe inventory systems in actual practice. Nor have we considered in detail problems involved in implementing any theoretical models. We feel, however, that a brief discussion here of implementation problems will help the reader to view the theoretical results in proper perspective. This part of the study, then, will indicate the types of practical problems that one can expect to encounter.

Perhaps the greatest difficulty in analyzing real-life systems or in installing new systems is obtaining the necessary data. The various costs associated with operating the system are often difficult to estimate even approximately. While it is to be expected that stockout costs would be difficult to quantify, trouble also develops where one does not foresee complications: the cost of placing an order, for instance, turns out to be extremely difficult to determine. In addition to this difficulty with costs, there are usually problems that prevent our learning much about the stochastic processes associated with the inventory system. One will often find it very hard to become well-informed on the processes generating demands. This problem has been especially critical in devising ways to control inventories of low-demand spare parts in the military and in industry. Similarly,
little will be known about the processes generating procurement lead-times (often less than what is known about the processes generating the demands). The difficulties just mentioned lie not only in the fact that the information is seldom available when one wishes to study the system, but also in the fact that it is extremely difficult to generate even if the necessary time is available.

Another problem which can hardly be avoided in real-world systems is human error. Even if the best model imaginable were available, and all necessary data were known, the resulting system could easily be rendered useless by the people operating the system. The model is not subject in this way to human error. Let us examine an example. One large department store introduced a hybrid (Q, s) system with periodic review. The greatest problems encountered in operating the system resulted from sales persons' making incorrect inventory counts, or from clerks' not ordering or ordering the wrong quantities. (Similar mistakes had been made under the old system, and the introduction of the new system served to focus attention on them even if it did not itself correct them.) Human error can of course be reduced if employees are carefully screened and if they have concrete incentives to do their jobs well. But even then, mistakes can occur.

In the real world, evaluation of an inventory system often turns out to be a tricky matter. In theory, the evaluation problem is simple: all relevant considerations are assigned a cost, and the optimal system is the one which minimizes the total cost. In practice, however, where a complex system can be subjected to a wide variety of stresses and the system can interact with its environment, changing both, it is often hard to
say what is even meant by optimal, and the problem of evaluation becomes a very real one. Often when one questions businessmen about their adoption of, say, a lot-size, reorder-point system, they can list certain advantages of the new system and they know what it cost to install it, but they are unable really to answer the question as to whether the new system is better than the old one or not. Part of the reason, of course, is that it is no easy task to isolate cause and effect. For example, profit on a particular item may go up after installation of the new system, but there is not necessarily any direct relation between this and the system introduced. Other economic factors, other things taking place within the firm, etc. can also exert influence. Another reason is that there is usually no single standard by which the results are judged. The new system may have reduced inventories, but the buyers may now be unhappy because (under the new system) they have less control over the buying operation.

A different facet of the evaluation problem involves the following. It may turn out that one of two systems seems to have a clear advantage over the other when the system is subjected to one set of stresses, while precisely the reverse will be true when a different set of stresses is applied. Both sets of stresses may occur in the real world, and the question remains how one would determine which system is to be preferred. Or perhaps one system will work very well when everyone does his job properly but it will deteriorate rapidly when the efficiency of the people involved goes down, while another system may not be quite so good when everyone is at peak efficiency, but will not deteriorate nearly so rapidly when efficiency declines.
In theory, as mentioned above, one would simply assign a cost to each of the attributes mentioned, and there would be no problem, provided that the optimization could be carried out. In practice, however, one can usually look only at real out-of-pocket costs, and these in themselves are not sufficient to evaluate the behavior of a system. Many other factors enter in, so that it often becomes a very difficult task to evaluate a system.

A type of problem distinct from those referred to above can sometimes occur. An available system which would do a good job in controlling the inventories may simply be too expensive to use. For example, the dynamic programming model referred to earlier might be quite suitable, say, for controlling the women's fashion goods of a department store, but could not be used because a large digital computer would be required for the fashion goods department alone. In other words, there are sometimes situations where superior control schemes are available, but cannot be used because their cost is not justified by the savings that would result from their use.

Up to this point, discussion has centered on problems which can arise even if an entirely adequate model is available for installation as a control scheme. Now problems involved with the models themselves will be examined.

The type of model that has appeared to receive the most attention from the standpoint of attempted application is the lot-size, reorder-point type with continuous information concerning stock levels. Attempts to apply this model to military use seem in some cases to have been premature. The model has been tried on situations where the
demand was far from stationary. Other attempts have run into difficulties in the form of budget constraints which made the transition phase impossible. Various ad hoc measures have been devised to retain the model despite such budgetary constraint, but none has proved adequate. In addition, the effects of redistribution costs on the inventory models have not been included in a satisfactory manner.

Other criticism of attempted military applications for the lot-size, reorder-point model stems from the fact that the model is designed for a single-echelon supply system. As pointed out above, the effects of applying this model at one echelon or another echelon or echelons, or to the system as a whole, have not yet been analyzed. Without such analysis, it is by no means certain that the model is appropriate for installation in a multi-echelon system.

Quite a number of applications of lot-size, reorder-point models have been carried out, both here and abroad. Among the companies that have utilized these models are General Foods Corporation, Westinghouse Electric Corporation, Crane Company, Edison Volta, J. L. Hudson Company, Famous Barr, Gimbel's, J. W. Robinson Company, International Business Machines Company, and Radio Corporation of America. Some of these applications have apparently been quite successful, while others have not achieved the desired objectives. One prominent department store, for instance, found that the only observable effect was an increase in inventories while sales levels remained constant. Other firms were able to decrease inventory levels and stockouts substantially. In several instances, it appears that the savings were, for the most part, brought about simply by having any system at all rather than chaos.
Gross overstocking of some items was discovered, along with understocking of others. No published analyses providing a detailed evaluation of these applications seem to be available. Some firms had not commented on the results beyond expressing vague hopes that the system was working well.

Of course there are factors at work preventing objective, detailed analyses. Most of the systems are installed by consulting firms which naturally cannot be expected to be completely objective concerning the results. Firm personnel who are responsible for hiring the consulting services also may find it hard to be objective. Furthermore, if the system is a tremendous success, a firm will not usually be anxious to inform its competitors about it.

Another type of model that has received considerable attention from the standpoint of military application is the simple one-period model (described earlier) used to calculate optimal "flyaway kits," i.e., spare parts to be included in a package of given total weight, designed so that the expected number of shortages of parts will be minimized over a specified period of time. Application of a similar type of model is being attempted for determining the range and depth of stockage of spare parts for a submarine on a cruise of specified duration, where constraints are placed on space and budget. Two of the principal difficulties in these applications are that the demand rates are, to a large degree, unknown and that stockout costs are difficult to establish. These factors of uncertainty are not necessarily reasons for not applying the model, since elements of unpredictability exist in any event. Obviously, it is not easy to evaluate the success of the models which have actually been applied.
A final example will illustrate the way in which the theoretical models are often modified (appropriately or not) for practical application. Recently, a management consulting firm applied in the retailing area a pseudo lot-size, reorder-point model with discrete review intervals. The application was to the cosmetics department of a retail store with several branch outlets and a central warehouse. The consultants did not provide the store with any theoretical justification of the particular model used. Significantly, the management was totally uninterested in any theoretical justification, and not especially interested in detailed analysis of the results. The warehouse ordering quantities were set by the use of the simple lot-size formula, \( Q = \sqrt{\frac{2\lambda A}{IC}} \). Inventory items were divided into three categories, A, B, and C. The A group consisted of 10 per cent of the items, which accounted for 50 per cent of the sales; the B group consisted of 40 per cent of the items, accounting for 40 per cent of the sales; and the C group the remaining 50 per cent of the items, accounting for only 10 per cent of the sales. The A, B, and C items were reviewed every 14, 26, and 56 days, respectively. If, at the time of review, the level of onhand inventory was below a level \( k \), an order was placed for a multiple of \( Q \) which would bring the onhand plus on order quantity above \( k \). It remains only to describe the setting of the reorder point levels, \( k \). These levels were set by Poisson tables, the largest value of \( k \) being selected, such that

\[
P\left(k, \lambda \left(L + \frac{R}{2}\right)\right) \geq .05,
\]

where \( P\left(k, \lambda \left(L + \frac{R}{2}\right)\right) \) represents the Poisson probability that demand
will exceed \( k \) in time \( L + \frac{R}{2} \), given that \( L \) is the leadtime and \( R \) is the review interval, both in terms of years.

At the individual stores, the inventory items are reviewed on the same schedule as for the warehouse. The policy at these stores is to order up to a maximum level, \( M \), at each review. The level \( M \) is the highest value of \( M \) for which:

\[
P \left[ M, \lambda(L + R) \right] \geq 0.05,
\]

where \( P \left[ M, \lambda(L + R) \right] \) represents the Poisson probability that demand will exceed \( M \) in time \( L + R \). The leadtime, \( R \), is assumed to be 5 days except for the store located where the central warehouse is; there the leadtime is only one day. Inventories at the centrally-located store are also reviewed more frequently than at the branch stores, a weekly review being usual.

Although no tangible savings could yet be attributed to the program, it was alleged by the inventory control manager that the results of the system were faster and more accurate reordering, faster and more efficient decisions, additional sales analysis, and more time for the buyers to make important decisions instead of worrying about operational details. Some additional clerical help was required.

CONCLUDING NOTE

The review of unsolved theoretical problems and of the difficulties of applying existing inventory theory in practice serves to bring out some of the challenging tasks remaining for the theoretician and the practitioner. The first will be concerned with the development of more refined models that can satisfy practical situations. The second will
have to make procedures devised by theory compatible with the concrete organization. The practical benefits to be derived from the theoretical efforts will depend strongly on the success of both the theoretician and the practitioner in these pursuits.
REFERENCES


10. Morse, P. M., Queues, Inventories, and Maintenance, John Wiley and Sons, New York, 1958, Chap. 10.


