Issued Under Office of Naval Research
Contract no. Nonr-1858(16)

PRODUCTION STABILITY AND INVENTORY VARIATION

Production Change Costs and Their Effect Upon Inventory and Production Policy

Daniel Orr

Econometric Research Program
Research Memorandum no. 15
13 May 1960

Princeton University
Econometric Research Program
92-A Nassau Street
Princeton, N. J.

ERRATA

Page	Line	
3	5	Read: $\geq$ for $>$
8	10	"specified"
9	13,15,16	Read: $L(X_i,Z)$ for $J(X_1,Z)$
12	footnote	Title incompletely italicized
16	footnote	"Elementary"
30	4	Read: (29) for (28)
37	22	Read: $(k + 2r)\ (\vec{\mu} + \underline{\mu})$

PREFACE

1 Background

This study is one outgrowth of an assignment undertaken in the Fall of 1956 at the Industrial Engineering Division of the Procter and Gamble Company. Another outcome of the same project is the "inventory controlled scheduling system" now in use or being installed at the Company's several domestic manufacturing facilities. The concern of that assignment was to find a scheduling policy which took into account the costs considered in most of the scheduling analyses then available, but which at the same time incorporated a feature which was viewed as critically important by manufacturing management: the possibility that large fixed costs may be incurred as a result of changing the production rate.

Production change costs have been incorporated into several analyses of inventory and production processes. These costs are usually formulated as linear functions of the magnitude of interperiod change. For example, consider the policy

$$(0) \qquad z_t = \hat{\xi}_t + \pi(X^* - X_{t-1})$$

where z_t is scheduled production, $\hat{\xi}_t$ is the forecast of demand, π is a smoothing constant $(0 < \pi < 1)$, X^* is an ideal or target inventory, and X_{t-1} is inventory on hand at the beginning of period t. This policy is suitable for controlling costs of production change which depend only upon the size of the change: by decreasing the constant π, output can be smoothed. However, the existence of a production change was held to be a critical factor in determining the costs incurred, so that control of

the size of the change is not all that the firm's interests require. The number of changes must also be kept at a reasonable level. A production policy which takes this objection into account is the subject of this study.

The investigation of this policy is divided roughly into two major phases: the first is concerned with determining by analytic methods optimal values for the decision variables and parameters specified by the policy; the second is an investigation of the factors responsible for production change costs, with emphasis on the way the new policy is able to control these costs, in addition to the costs normally considered in formulating production policies. Chapter II is devoted to the first phase, while the second is completed in Chapter III. From the viewpoint of traditional economic analysis, this second phase investigation is the more interesting. Ideas borrowed from standard work on the theory of production serve as a basis for the discussion of dynamic production costs.

The first phase of the investigation is similar to previous work in the mathematical theory of inventory and production. Recent studies of production and inventory processes have used analytic methods similar to those first employed by Abraham Wald in his work on sequential analysis. The present study relies even more heavily on this work than its predecessors, since the policy here investigated has a form which makes it expedient to utilize Wald's results directly in developing approximation solutions.

Chapter I has four main purposes: the first is to describe the problems encountered in dealing with the type of production change cost visualized in this study; the second is to describe two rigorous analytic methods (the recursive functional equation approach of dynamic programming, and the standard method of obtaining the stationary distribution of a Markovian inventory process) which have been successfully used in conjunction with other inven-

tory policies and which have been considered as possible approaches to the present problem; the third is to describe certain approximation techniques which have served to simplify the computational problems frequently encountered in dealing with inventory processes; and the fourth is to review that portion of earlier work in which different formulations of production change cost appear.

In Chapter II the new production policy is described, and in an appendix the attempts to analyze this policy by rigorous analytic methods are reproduced. The approximations based upon sequential analysis are developed and suggested as an alternative when more rigorous analysis is not successful.

Chapter III is an investigation of production cost and production change cost. Support is developed for the contention that production change costs will depend (a) upon the frequency of production changes, and (b) upon the number of different production rates specified by the policy.

Chapter IV is concerned with methods for determining when the new policy is suitable for use; several important questions regarding its form are answered. It is also appraised from the standpoint of application.

2 Acknowledgements

At Procter and Gamble, principal sources of stimulation and help were George Montillon, Cornelius Muije and Ram Gnanadesikan; the latter guided me through the literature of mathematical statistics, and discussions with the first two generated numerous ideas. Associates at the Econometric Research Program and at Mathematica: Michel Balinski, Arlington Fink, Clive Granger, Irwin Guttman, and Roger Pinkham, were helpful at several stages of the study. Professors Harold Kuhn and William Baumol made numerous valuable suggestions. The greatest debt is to Professor Oskar Morgenstern, director of the Econometric Research Program, where this study was done under contract with the

Office of Naval Research. His advice, encouragement and support are gratefully remembered. Mrs. Patricia Granger did most of the typing of a sometimes perplexing manuscript. My wife Mary Lee was helpful, compiling references and assisting in editing, proof reading, and typing.

C O N T E N T S

CHAPTER I

SCOPE AND METHOD

1 Introduction

The fundamental idea of this study is that in a wide variety of manufacturing processes, changes in the output rate can be effected only at some substantial cost, even when these changes are very small in size. Clearly, the only way to control such costs is to control the frequency of production rate changes. This may be done by specifying that until inventory accumulates to some prearranged high level, the rate of production will not be decreased; and until some low inventory level is reached, production will not be increased. Thus, a range of inventory levels is identified, within which production is maintained at a constant rate. This is the germ of a policy for controlling production; however, the vague description just recorded must be made more concrete. Specific numbers must be assigned to the high and low inventory levels, and methods must be given for determining how large production changes should be when they are made. Mathematical methods recently have been successfully applied to such problems; this approach will be taken in the present study.

A production and inventory system is abstractly represented as consisting of four major parts: a sequence of future demands

(1) $\xi_1, \xi_2, \cdots,$

a replenishment policy

(2) $Z(X_0; \xi_1, \xi_2, \cdots),$

an objective function

(3) $L[Z(\cdot)]$,

and a material balance relation

(4) $X_t = X_{t-1} + z_{t-\lambda} - \xi_t$;

X_t is inventory at the end of period t (at the beginning of period t+1); and z_t and ξ_t are the quantities of replenishment arriving in inventory and demand taken from inventory during the period t. The time lag $\lambda \geq 0$ is a delay in the flow of material or information between the placement and the arrival of a replenishment order.

The demand series (1) is abstracted in a variety of ways: it may be treated as deterministic, with exact knowledge as to future timing of demands and amount demanded; or stochastic, stationary and independent, with known distribution; or any of the simplifying qualifications on stochastic demand may be dropped, and the only available information regarding demand may be the time series of orders generated since the inception of the firm. (In this study, attention will be confined to models which assume stochastic demands.) In dealing with stochastic demands, it is convenient to view time as a sequence of equal discrete intervals, e.g. weeks, with decisions taken once each period, perhaps at the beginning. A decision in period t is evaluated on the basis of the cost it incurs in period t, but also on the basis of future costs which it is likely to incur.[1] The number of future periods considered in making present decisions is called the planning horizon; in certain instances it has been found expeditious to specify an infinite horizon, i.e. to discount costs over an infinite number of future periods.

The policy is the sequence of decisions over the entire planning

---

[1] The future impact of a present decision may be traced through the balance identity (4): next period's starting inventory is affected by the replenishment decision in this period.

horizon. When the policy (2) consists of a simple set of rules which define a response to all possible contingencies, it is called a policy of simple form.[2] An example of such a policy is: If inventory is above the level X^*, order nothing; if inventory is below X^*, order up to X^*.[3] Or more concisely,

$$(5) \quad z_t = \begin{cases} X^* - X_{t-1} & X_{t-1} < X^* \\ 0 & X_{t-1} > X^*. \end{cases}$$

Like most policies of simple form, this policy identifies one or more critical inventory levels, and the replenishment volume is regulated according to the position of actual inventory vis-a-vis these critical levels. (The policy (0) described in the preface is one example in which factors other than the level of inventory are important in determining the replenishment volume.)

A variety of cost structures can be visualized, with differences arising in different physical situations. Most analyses recognize costs of holding inventory, costs of running out of inventory, and costs associated with replenishing inventory. Two different types of inventory system with markedly different cost structures may conveniently be identified: ordering systems, in which replenishment is from an outside source; and production systems, in which replenishment is from within the same organization. Different costs of replenishment will be encountered in the two types of system.

---

[2] Cf. Kenneth J. Arrow, Samuel Karlin and Herbert Scarf, Studies in the Mathematical Theory of Inventory and Production (Stanford, California: Stanford University Press, 1958), p. 223.

[3] This particular policy is analyzed at length by Richard Bellman: see his Dynamic Programming (Princeton: Princeton University Press, 1957), pp. 152- 182.

Having identified the costs of the objective function, the most ambitious objective is global optimization, the selection of the optimal policy from the infinite set of possible policies, where cost minimization is the criterion of optimality. Due to computational difficulties, this objective proves feasible only for extremely simple models. Frequently, the form of the policy is specified before undertaking cost minimization, and the problem is to find the best of the subset of admissable policies having the specified form. For instance, it may be determined in advance that a policy of the form (5) will be used; the problem then is to find the optimal value for the policy parameter X*.

Frequently, physical constraints such as production or inventory capacity are most conveniently incorporated into the model by imbedding them in the objective function: for instance, an inventory capacity might be represented by an abrupt increase in the inventory cost function at the capacity level, since storage alternatives become more costly at that level. As another example, the factor which distinguishes the present study from previous work in inventory and production analysis is the attribution of cost to the act of changing the production rate. These costs are incorporated into the loss function as fixed costs of making production changes. The act of making a change in production is assigned a substantial cost, independent of the size of the change, and the frequency of changes is thereby controlled. Although this representation enables incorporation of this type of change cost into the loss function, it leads to serious computational difficulties which are a well-known concomitant of fixed charges. A substantial discontinuity is encountered in going from changes of size zero to changes of size one, so the plot of cost as a function of the size of change is non-convex, and hence may have more than a single local minimum.

The function which plots cost against size of change will differ for each possible initial inventory level X_0. For example, when X_0 is neither too high nor too low, the function will look somewhat as in Diagram I-A. There the best course of action is not to change. However, when X_0 is too low, an increase may be indicated, as in Diagram I-B; or when X_0 is too high a decrease is called for, as in Diagram I-C. Each of these functions has a local minimum where the size of change is zero, and therefore standard optimization techniques are unreliable.[4]

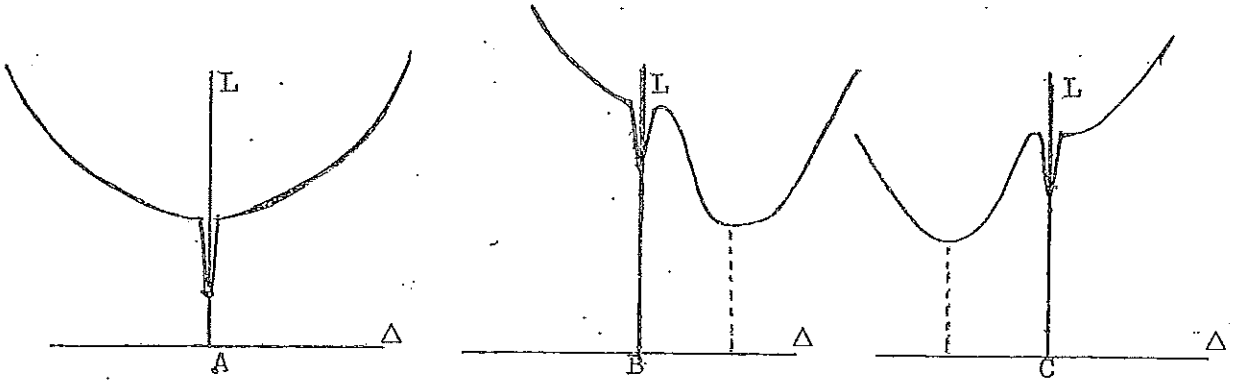

Diagram I

Total Cost L as a Function of Change Size Δ For Different X_0.

In the process of formulating the loss function, i.e. describing the response of costs to various properties of the sequence of production, inventory and sales, two different approaches have been taken. The first, called the smoothing approach, employs statistical properties (e.g. means and variances) of the sequences of production and inventory as measures of the expected long-run cost resulting from a particular policy. The second, called the transient, or dynamic approach, is concerned with more specific measures, such as the probability distribution of inventory levels one period hence, or the probability of not meeting some prespecified percentage of customer demands over the next three periods. Rigorous analytic techniques

[4] For further discussion of difficulties caused by discontinuity and non-convexity in the objective function, see William J. Baumol, Economic Theory and Operations Analysis (Englewood Cliffs, N.J.: Prentice-Hall, forthcoming in 1960), Chapter Six.

are associated with each of these approaches; these are described next.

1.1 Stationary Analysis

The most elegant and rigorous analytic formulation which takes the smoothing approach is the "Steady State Solutions" of Karlin.[5] (Similar in intent are the "servomechanical" analyses which will be described in Section 2 of this chapter.) Stationary (steady state) analysis is based upon the following argument. Suppose present inventory is X_0. The material balance identity (4) indicates this fact is critical in determining expected inventory one period hence, X_1, but looking farther and farther into the future, the effect of X_0 on the successive inventory positions becomes less and less pronounced; until in the limit, predictions of average inventory levels, or probabilities of being between two specific inventory levels, are not less accurate if information regarding the present inventory level is ignored entirely. More precisely, the stationary approach defines a transformation T and the relation $X_n = T^{(n)}X_0$;[6] its central hypothesis is that the process is convergent, i.e. that the probability density is given by

(6) $f(X)dX = P(X \leq \lim_{n\to\infty} T^{(n)}X_0 \leq X + dX)$.

The distribution function is written $F(X)$.

The transformation T in any period is determined by two factors; the first is the quantity demanded in that period, and the second is the procurement response, as determined by the policy (2). It is usually safe to assume that if the demand series is stationary, and the policy

---

[5] Arrow, Karlin and Scarf, op. cit. pp. 223-269.

[6] This notation indicates that the transformation T is applied to the initial state X_0 n times in succession.

operates in a way such that high inventory is countered by reduced procurement and vice versa, then the inventory distribution will be stationary, i.e. the convergence property (6) is likely to hold. A rigorous analysis might furnish a complete proof of this convergence; other analyses might prove only that F(X) has finite mean and variance; still others might proceed as though stationarity were assured, using derived properties of the stationary distribution in an optimization routine, without formal justification. Regardless of how they are obtained, properties of the stationary inventory distribution can be related to the expected future pattern of costs in the system. Then the effect of the policy on cost can be studied through its effect upon the stationary operating characteristics of the production-inventory-demand complex.

Karlin analyzes the stationary characteristics of several simple policies. His proofs of stationarity, even in the simple case of Markovian processes, rely upon advanced topological arguments. It will be seen in the Appendix to Chapter II that this method encounters substantial computational difficulties when applied to policies which are more complicated than the ones he analyzes.

1.2 Dynamic Solutions

The inventory process described by equations (1-4) is an example of a general class of dynamic decision processes, which may be characterized as follows: A dynamic process consists of a sequence of states $(X_0, X_1, \cdots, X_{p-1})$, a sequence of decisions $(z_1, z_2, \cdots, z_p)$ and a sequence of autonomous events $(\xi_1, \xi_2, \cdots, \xi_p)$ where p is the number of stages (time periods) in the planning horizon. The successive states are tied together by a relation called the state transformation function, which in our abstract notation is written $X_{i+1} = z_i(\xi_i, X_i)$; in the inventory

problem this function is the balance relation (4). Characteristically, this function is determined by the decision (e.g. the amount of replenishment) and the autonomous factor (the amount demanded by customers). Associated with each possible combination of decision and state at every stage of the process is an expected payoff, or return, written $E\,[R_i(X_{i-1}, z_i)]$ to emphasize the dependence of the expected return upon the initial state and the decision. The complete sequence of decisions over the planning horizon is the policy (2), and the discounted sum of the sequence of payoffs over the entire horizon is the objective function (3). The dynamic programming problem, then, is to select a completely sepecified policy from either the set of all policies or a subset of partially specified policies, in such a way that the objective function is maximized, i.e.

$$(7)\quad L(X_0, Z) = \max_{z_i} \left\{ E\left[\sum_{i=1}^{p} d^{i-1} R_i\,(X_{i-1}, z_i)\right] \;\middle|\; z_i \in Z,\ X_i = X_{i-1} + z_i - \xi_i \right\},$$

where d is a discount factor.

If the transformation $z_i\ (\xi_i, X_i)$ has a stochastic component (as is the case in the inventory problem when demands are not known with certainty in advance) we assume probability measures are defined, and maximize the expected value of the discounted future stream of returns, as shown. The problem of solving such a multidimensional relation containing a maximum operator may be formidable. A fundamental lemma has been known for some time,[7] which Bellman calls "the principle of optimality": it is stated

> A policy which is optimal over the entire horizon period is optimal starting at any intermediate point and operating to the end of the horizon. [8]

---

[7] This principle is recognized in an early study: E.S. Shaw, "Elements of a Theory of Inventory" Journal of Political Economy. Vol. XVIII, 1940, pp. 465-485.

[8] Bellman, op.cit. p.83.

This lemma has been the key to the development of solutions for dynamic programming problems.

In the inventory problem, the particular value of the principle of optimality has been in the development of analytic optima for continuous, stochastic invariant processes. For example, suppose the planning horizon p is infinite, the discount factor d is constant, and the distribution of customer demands $\Phi(\xi)$ is stationary. Then the recursive structure of the process may be utilized to rewrite the relation (7) as

$$(8)\quad L(X_0,Z) = \max_{z_1}\left\{ E[R(X_0,z_1) + d\ L(X_1,Z)] \,\middle|\, z_i \epsilon Z,\ X_i = X_{i-1} + z_i - \xi_i \right\}.$$

Equation (7), which was a function of an infinite number of decision variables, has been transformed into a functional equation calling for maximization over a single decision. Equation (8) has the property that if Z is the optimal policy, and the initial state X_0 is not specified, then $J(X_0,Z)$ and $J(X_1,Z)$ are equivalent: both represent the maximum future earnings stream from an unspecifed initial state, discounted over an infinite horizon. Stated another way, the transformation of $J(X_1,Z)$ into $J(X_0,Z)$, effected by the decision z_1, is a mapping of the maximum return function into itself. Thus the policy Z we seek is the one associated with a fixed point of the functional relation (8). Thus the problem is reduced to solving for a fixed point, which may still be no easy task. In several notable instances,[9] however, formulation of a problem in the invariant recursive form (8) has led directly to solution. As

[9] Richard Bellman, Irving Glicksberg, and Oliver Gross, "The Optimal Inventory Equation," Management Science, Vol. II, 1955, pp. 88-104 is an example of global optimization for an extremely simple ordering problem by dynamic programming, as is Karlin (Arrow, Karlin and Scarf, op. cit. pp. 155-178.) Most notable and earliest application under the constraint that the form of the policy is prespecified is Kenneth J. Arrow, Theodore E. Harris, and Jacob Marschak, "Optimal Inventory Policy" Econometrica, Vol. XIX, 1951, pp. 250-272.

in the case of steady-state solution, the amount of complexity which may successfully be incorporated into these formulations is limited.

1.3 An Approximation to Stationary Analysis

In the development of inventory theory, rigorous analyses of the sort described in Sections 1.1 and 1.2 are a comparatively recent development. Prior to the 1951 study of Arrow, Harris, and Marschak,[10] the literature on this topic all relied upon an approximation formulation which avoids the computational difficulties inherent in the more rigorous formulations, without distorting the model to the extent that it is a complete misrepresentation of any physical situation. Typical of this approximation approach is the work discribed and extended by Whitin.[11] The analysis of the production policy proposed in this study relies upon such an approximation approach.

The rationale underlying a rigorous stationary analysis has already been explored. Briefly summarizing, the aim is to describe explicit relations between the costs of an inventory system, and critical probability levels of the stationary inventory distribution. For example, suppose the s,S policy is used. This policy is described as follows:

> Establish two critical inventory levels, S and s, $S > s > 0$, with the operating rule: when inventory falls below s, order enough to bring the level up to S; when inventory is above s, order nothing.

---

[10] Op. Cit. Actually, the priority we assign to this study is valid only in effect: Pierre Massé's work, Les Réserves et la Régulation de l'Avenir dans la Vie Economique, (Paris: Hermann et Cie., 1946, Vol. II) was not known in this country until after the publication of the paper of Harris et. al.

[11] Thomson M. Whitin, The Theory of Inventory Management, (Princeton: Princeton University Press, Second Edition, 1957) pp. 30-79.

Stated more concisely,

$$z_t = \begin{cases} S - X_{t-1} & X_{t-1} \leq s \\ 0 & X_{t-1} > s \end{cases} .$$

The expected cost per period of holding inventory is readily related to the stationary inventory distribution: it is simply the product of cost times probability of each inventory level, or

$$(9) \quad \int_{-\infty}^{\infty} \eta(X) f(X) dX$$

where $\eta(X)$ is the holding cost function, and $f(X)dX$ is the stationary inventory density. The expected cost per period of ordering is obtained by multiplying the cost per order by the probability of being below the reorder point s,

$$(10) \quad K \int_{-\infty}^{s} f(X) dX$$

where K is the cost of placing an order, and ordering costs are assumed independent of the order size. The cost of shortages is given by multiplying the probability of being below zero inventory (or in instances where unfilled orders are not backlogged, at zero inventory) by the penalty incurred each time the system is out of stock at the end of a period, which is assumed to be a lump-sum cost:

$$(11) \quad \rho \int_{-\infty}^{0} f(X) dX .$$

Since each of these costs is readily related to the stationary inventory distribution, and since the distribution depends upon the decision parameters (the maximum inventory S and the reorder point s) this approach makes it possible to study the impact of different pairs of these values upon expected total cost.

The stationary analysis of the s,S policy utilizes concepts which depend upon the assumption that time consists of a sequence of discrete

intervals. In particular, the structure usually envisioned for demands (that they are a sequence of independent, identically distributed random variables) is impossible without this assumption. The approximation approach which Whitin uses is based upon an asymptotic case in which these time intervals all approach zero in duration. Unless there is a time lag in delivery [$\lambda > 0$ in equation (4)] the necessity to set s, the reorder point, above zero arises only because inventory might fall below s between checks upon the state of the system. As the interval between these checks approaches zero, so does the level s, the inventory required to protect against depletion between checks.

The approximation method [12] calls for what is in effect a continuous surveillance of the inventory system; thus when $\lambda = 0$ in (4), the only costs involved in the decision are holding cost and ordering cost. In using this approach, holding cost is formulated as a linear (or other more general) function of the average inventory level. If drains upon inventory take place at a constant rate through time, the average inventory level will be $S/2$, so (9) is written $\eta(S/2)$, or when holding costs are linear,

(12) $\eta\, S/2$.

In a similar fashion, expected values are used to approximate the costs of ordering. As an analytic convenience, the planning horizon is arbitrarily set at one year. Then, the expected number of orders per year is approximated by $E(\xi_y)/S$, where the numerator is the mean annual demand. The

[12] The mathematical basis of this approximation analysis is investigated by Donald M. Roberts; "Approximations to Optimal Policies in a Dynamic Inventory Model" (Applied Mathematics and Statistics Laboratory, Stanford University, Stanford, Calif.: Technical Report no. 12, July, 1959).

expected ordering cost, then, is represented by

(13) $ME(\xi_y)/S$.

The sum of (12) and (13) may be taken as a representation of those costs which vary with the replenishment quantity. It was recognized some time ago[13] that the derivative of the resulting sum with respect to S gives the necessary condition for optimality of the order quantity. This result, which is derived by elementary calculus, is offered in lieu of the more circuitous stationary analysis. In some instances, the assumption that inventory levels are under constant surveillance, which is used in the approximation approach, may be a more accurate representation of the existing control practice than the once-per-week review assumption of the more formal analysis. Further, the approximation formulation requires less information about customer demands, using only the mean per annum, instead of the distribution per period. If the assumption that the time lag $\lambda = 0$ is relaxed, the approximation approach handles this complication nicely by incorporating a positive reorder point in the analysis.[14] The approximation policy which deals with lags in delivery is known as the two-bin policy.

2. Stochastic Models of Inventory and Production

The principal motive of this study is the analysis and accommodation of production dependent costs, so a review and summary of those analyses which have previously concerned themselves with these costs is indicated.

[13] This result was first proposed in 1925; Whitin op. cit. cites three authors who published it in that year.

[14] This is handled by Whitin, op. cit. Chapter 3. Compare his simple analytic methods with the more sophisticated (but perhaps less effective) tools used by Karlin and Scarf (Arrow, Karlin and Scarf: op. cit. Chapter 10.)

The bulk of the literature of inventory theory has been concerned with problems of ordering, rather than production;[15] however, this distinction has not always been clear-cut; for example, Whitin suggests the two-bin policy as a practical program for production control.[16] The general validity of such an approach is to be doubted: first, if only one product is manufactured at a given production facility, it is likely that the procurement policy will emphasize smoothness of operation, rather than size of runs.[17] Conversely, if more than one product is cycled on the same production equipment, and the economical lot run size is calculated independently for each product, there is no assurance that the total schedule will be feasible; the lot run of product A may be completely exhausted before the runs of B, C, ... have been completed. More sophisicated smoothing analyses of the single product problem are available; some of these are described next.

Simon offers an interesting resolution of the computational difficulties inherent in the dynamic programming approach.[18] He proves that if all costs in the model are quadratic functions of their arguments, inventory

---

[15] This is true of most of the work already cited: Bellman, Arrow, Harris and Marschak; Whitin, and the bulk of Arrow, Karlin and Scarf. A portion of this last book is devoted to production problems, but in dealing with these problems demands are assumed known in advance.

[16] Whitin, op. cit. Chapter 3, especially p. 33.

[17] Except perhaps for certain batch process industries (e.g. oil refineries) which are operating at substantially less than capacity. (If they were enjoying capacity operation, physical capacity constraints would presumably be the primary factor in determining economical lot run size.)

[18] Herbert Simon, "Dynamic Programming Under Uncertainty With a Quadratic Criterion Function" Econometrica, Vol. XXIV, 1956, pp. 74-81.

and production, then a stochastic dynamic programming problem may be transformed into a deterministic one, using the expected values of stochastic variables as "certainty equivalents" (single-valued predictors). Although the objective function may no longer be written in a general way, this is partially compensated by more flexible allowable assumptions regarding future demands: these need not be assumed stationary and independent. The following costs are assumed in Simon's analysis: (1) inventory costs – costs associated with different positive and negative levels of inventory, i.e. holding cost and runout penalty in our previous terminology, written $C_i\ (X_t - I_c)^2$, where C_i and I_c are a cost and inventory parameter respectively; (2) costs associated with different levels of the static production function – $C_p(z_t - P_c)^2$, where C_p and P_c are a cost and a production parameter respectively; (3) costs associated with magnitude of change in the production rate – $C_{pa}(z_t - z_{t-1})^2$, with C_{pa} a cost parameter. It is desired that the function comprising the cost terms and the balance identity (4)

$$E[L(X_0)] = C_i \sum_{t=1}^{N} \left[\sum_{j=1}^{t} \xi_j + X_0 - I_c \right]^2 + C_p \sum_{t=1}^{N} (z_t - P_c)^2 + C_{pa} \sum_{t=1}^{N} (z_t - z_{t-1})^2$$

be minimized over the N-period planning horizon. He uses standard techniques from the calculus of variations to show that optimum production depends upon no property of demand except its expected value in each period of the planning horizon. Boundary value considerations, which typically obtrude in dynamic programming, are avoided here; there is strong incentive not to operate at boundaries of arguments due to quadratically increasing costs. Simon's model achieves essentially the same results as the smoothing analyses of Mills and Pinkham; this is an example of the ability to control the behavior of a system by regulating the objective function. In particular, the third cost term (which

penalizes changes in the period-to-period production rate) places a heavy premium on a smooth output of production decisions. These other analyses just mentioned do not deal explicitly with a loss function; it is left to the reader to bridge the gap between costs within the system and the statistical measures presented in the analysis.

Mills[19] avoids the problem of proving stationarity in his smoothing analysis by defining a control policy as stable, provided all relevant first and second moments approach finite limits, i.e.

$$\lim_{n\to\infty} E(X_n) = \bar{X}, \quad \lim_{n\to\infty} E(z_n) = \bar{z}, \quad \lim_{n\to\infty} E(\xi_n) = \bar{\xi},$$

$$\lim_{n\to\infty} E[X_n - E(X_n)]^2 = \sigma_x^2, \text{ etc.}$$

Given a stable process, ratios of interest are

$$k_X = \sigma_X^2/\sigma_\xi^2, \qquad k_z = \sigma_z^2/\sigma_\xi^2.$$

Using a linear combination of the ratios k_X and k_z as criterion, the optimal policy can be shown to belong to a simple policy class of the following form:

$$(14) \quad z_t = Az_{t-1} + (1-A)\,\xi_{t-1}, \qquad 0 < A < 1, \qquad \bar{z} = \bar{\xi}.$$

This class of policy establishes a lower bound on linear combinations of k_X and k_z, and by varying A, the linear combination consonant with cost minimization can be obtained. An increase in A sacrifices inventory stability for production stability, and vice versa. A necessary condition is that backorders must be permitted; in the instance where inventory runouts lead to lost sales, stronger conditions must be imposed, i.e. there must exist an upper bound B upon the demand distribution; the policy then is identical with (14) when z_{t-1} is less than B, and nothing

[19] Harlan D. Mills; Stochastic Properties of Elemantary Logistic Components Econometric Research Program, Princeton University; Research Memorandum no. 9, February, 1959.

is produced when z_{t-1} is greater than B. The demands per period are assumed to be independent and identically distributed, but the population distribution is not assumed known; the analysis relies upon sample statistics only.

The analytic tools of the servomechanism[20] engineer have been used to advantage in designing inventory systems using smoothing criteria. The early paper of Simon[21] which deals with servomechanical properties of inventory systems is interesting for two reasons: it furnishes a brief description of servomechanisms, and analyzes an inventory-production system in servo terms. The analytic content of the paper is of limited interest, since nowhere does the author come to grips with a realistically formulated load (demand sequence); the most sophisticated load he considers is a sinusoidal series. The policy analyzed is the one described in the preface. The usefulness of Laplace transforms is indicated when questions of stability and stationary behavior arise. Only the means for accommodating a suitably characterized demand (i.e. one which enables the inclusion of a random element, as well as a fluctuating deterministic element) is needed, to provide an eminently useful analytic apparatus. This deficiency is remedied in the paper of Pinkham.[22]

---

[20] The term servomechanism describes a wide variety of self-correcting automatic devices. Components are the load (demand in an inventory system), an output (the replenishment order) an input (the policy rules), and a correction of output in the direction indicated by the input (called feedback). For exposition at an elementary level, see H. H. Goode and Robert Machol, System Engineering, (New York: Mc Graw Hill, 1957) pp.456-480, or the reference of note 22.

[21] Herbert Simon, "On the Application of Servomechanism Theory in the Study of Production Control" Econometrica, Vol. XX, 1952, pp 247-268.

[22] Roger Pinkham, "An Approach to Linear Inventory-Production Rules" Operations Research, Vol. VI, 1958, pp. 185-189.

Pinkham considers the system described by

$$X_t = X_{t-1} + z_{t-1} - \xi_t$$

$$z_t = Az_{t-1} + BX_{t-1} + \hat{\xi}_t$$

where $\hat{\xi}_t$ is a forecast of demand for period t, A and B are constants, and demands are assumed stationary and independent. By use of the power series transforms of the two equations, A and B are determined so that cost is minimized. Cost is an increasing function of the variances

$$\text{Var } (z) = \lim_{n\to\infty} E(z_n - z_{n-1})^2$$

$$\text{Var } (X) = \lim_{n\to\infty} E(X_n^2)$$

where $E(X_n)$ is scaled to zero. He shows that his methods may be used to establish policies when an autocorrelated demand series is encountered, but does not analyze this case.

Each of the smoothing policies which have been briefly described shares a characteristic which is called linearity[23] by servomechanism engineers. Linear policies are easy to analyze (one can easily describe and control the operating characteristics) because they are amenable to transform methods. However, in many instances, a linear policy is not an apt device for controlling replenishment; this is true especially when fixed costs are encountered, such as the costs of ordering visualized in the s,S policy.[24] When such discontinuities in the cost structure are encountered, no standard method of analysis is foolproof, and analysis must proceed on the basis of what is discovered about the structural properties

[23] Linearity describes the relation of output to input: if the summed output responses to several inputs is the same as the single output induced by one input equal to the sum of the several inputs, the system is linear. A linear inventory policy is one which causes a replenishment change which is a linear homogeneous function of the inventory deviation from some ideal level.

[24] Supra. p. 10. Scarf recently proved that some s,S policy will be optimal whenever the costs of holding and runout together are convex, and the ordering cost contains a fixed component (i.e. is of the form a + bz, where a > 0 and

of the problem at hand. The s,S policy is a renewal process,[25] which has been the key to its analysis since the earliest discussion by Arrow, Harris and Marschak.[26]

In the remainder of this study the dynamic costs of production will be characterized in a way which may be descriptively appropriate in a wide variety of continuous process industries, and a scheduling policy will be proposed for use in situations where these costs are encountered. In this formulation, these costs contain a fixed component which makes a linear policy inappropriate. The policy here proposed generates essentially the same stochastic process as is encountered in sequential analysis. This structure is utilized to develop an approximation solution. The second chapter is an analysis of this new production policy: a model is described, a policy proposed, and dynamic programming and steady state analyses are attempted. Due to the computational difficulties encountered in these attempts, an approximation approach analogous to Whitin's version of the s,S policy is developed; Wald's analysis of the random walk of sequential analysis furnishes certain measures which are employed in this approximation.

---

[24] z is the quantity ordered.) Cf. Herbert Scarf, The Optimality of s,S Policies in the Dynamic Inventory Problem, Applied Mathematics and Statistics Laboratories, Stanford University, Stanford, California, Technical Report no. 11, April, 1959.

[25] A renewal process is a stochastic process which satisfies two conditions: the random movements never cancel each other (they are always in the same direction) and a cycle is identifiable within the structure of the process, so that it consists of a series of random departures from and returns to some fixed level. The standard reference is William Feller, "On The Integral Equation of Renewal Theory," Annals of Mathematical Statistics, Vol. XIII, 1941, pp. 243-267. It will be observed in the next chapter that the new policy proposed in this study is not a renewal process.

[26] Op. cit. p. 264.

CHAPTER II

ANALYSIS OF A POLICY FOR CONTINUOUS PROCESS PRODUCTION

1. Formulations and Analyses of Production Change Costs

As was seen in the first chapter, the idea that inventory analysis should take into account changes in the rate of production which occur in response to variation in the level of inventory is an idea of fairly long standing, but one which has never been completely developed. The models of Pinkham, Mills, and Simon were described; these implicitly or explicitly take into account production costs and production change costs. Each of these models nicely accomodates linear cost relationships, which may be used at least as approximations when all the elements of the objective function are smooth functions of their arguments, inventory X_t and production z_t. Thus, while the resulting policies all are specifically intended for production control, none are equipped to cope with cost discontinuities, e.g. fixed costs of production or fixed costs of changing the rate of production. In the Mills and Pinkham papers, inventory costs are expressed as a function of the inventory variance, and production costs as a function of the production variance. This is likely to be a poor representation in situations where fixed costs are important; for suppose that production change costs depend upon the number of changes made in the production rate over a planning horizon, as well as upon the size of these changes. Five changes of two units each have less of an effect on the production variance than one change of six units, so if the number of changes is consequential, the variance measure underestimates

total cost; on the other hand, if cost increases more than as the square of the size of the change (which is the relationship implied by use of the variance as a measure of performance), and fixed cost elements of the type just described are absent, a single deviation of six units may be more costly than ten deviations of two units. It may be desirable for many applications to establish alternative criteria which are not based exclusively upon these variances.

Occasional attempts have been made to devise scheduling rules which incorporate fixed costs of changing production, but a realistic formulation of these costs is typically sacrificed in order to obtain something that fits into a particular computational schema. This contention is supported by a comparison of two papers which recognize these costs explicitly and discuss them in some detail. Simon's formulation of total manufacturing cost includes an element he calls "....sticky costs, proportional to the rate of manufacture when that is constant, but not capable of being reduced immediately when the rate of manufacture declines";[1] Hoffman and Jacobs, on the other hand, consider production change costs ... which depend upon the magnitude of increase of production; they see output reductions as costless.[2] These two papers thus involve antipodal

[1] Simon: "The Application of Servomechanism Theory..." p. 265.

[2] Alan Hoffman and Walter Jacobs, "Smooth Patterns of Production" Management Science, Vol. 1, no 1, 1954, pp 86-94. This paper is motivated by a remark in Franco Modigliani and Franz E. Hohn, "Production Planning Over Time and the Nature of the Planning and Expectation Horizon," Econometrica 23, January 1955, p. 66n.: "In the course of this research some attention has been given to one aspect of costs which was not considered in the paper, namely the cost of changing the rate of production. This cost, which may be significant in certain concrete situations, has been neglected here because we did not find it possible to formulate some general and yet reasonable assumptions about the behavior of such costs, as we were able to do for other components." In this context perhaps "reasonable" is an euphemism for "computationally tractable".

descriptions of the same phenomenon.

2 A Model of Continuous Process Production

We turn now to the prime concern of this study: to formulate and analyze a policy which successfully copes with costs of changing production rates, where these costs are formulated in a way descriptive of a wide variety of physical circumstances. The model will employ the following assumptions:

1. One product is continuously processed in one factory.
2. Time is formulated as a series of equal discrete intervals. Demands in successive periods are independent, identically distributed random variables with known distribution; they are the only stochastic elements in the model.
3. Production rate changes are effected instantly [$\lambda = 0$ in equation (4)].
4. The production rate is the only decision variable.
5. The objective is minimization of total cost of inventory and production, which includes the following elements.

   a. A "comparative static" unit cost of production,

(15) $$m(z_t).$$

   This is the average total unit cost function of the theory of the firm.

   b. Dynamic costs of changing the production rate,

(16) $$\begin{cases} \bar{K} + \bar{k}\,(z_t - z_{t-1}) & z_t > z_{t-1} \\ 0 & z_t = z_{t-1} \\ \underline{K} + \underline{k}\,(z_{t-1} - z_t) & z_t < z_{t-1} \end{cases}$$

   c. Costs of holding inventory

(17) $$\int_0^\infty \eta\,(X)\,dF(X)$$

d. Runout costs, the penalty for failure to meet demands due to depletion of stocks

$$(18) \qquad \int_{-\infty}^{0} \rho(X)dF(X)$$

6. The stationary inventory distribution $F(X)$, which appears in the inventory cost expressions (17) and (18), is assumed to exist, but is not assumed to be known a priori.

2.1 A Policy to Control the Frequency of Production Changes

This section introduces a policy which accommodates production costs and production change costs of the type visualized in the model. This policy, called the (a,b,c) policy, relies upon the same straightforward rationale that makes the s,S or two-bin policies such appealing devices for coping with the simpler problems presented by fixed costs of ordering. These policies both counter "lump-sum" ordering costs (10) by controlling the frequency of orders. The (a,b,c) policy proposes that lump-sum costs of changing the rate of production (16) be countered by controlling the frequency of these rate changes. In the s,S and two-bin policies, an order is placed only when the expected cost of shortage outweighs the known cost of ordering; in the (a,b,c) policy, production is maintained at the minimum cost rate until a reduction in finished goods inventory causes sufficient concern over the expected cost of shortage to warrant an output increase, or until accumulation of inventory reaches a point where the cost of holding dominates the cost of decreasing the rate. At these times of low or high inventory, the production rate is increased or decreased, the change being maintained until inventory passes a predetermined level at which output is restored to the minimum cost rate. The policy is defined:

Establish three inventory levels, $a > c > b$, and three production rates $\bar{p} > p > \underline{p}$, with the operating rules: (a) When inventory X_t is between a and b, and production in this period $z_t = p$, set production next period, $z_{t+1} = p$. (b) When inventory accumulates to or above the level a, and $z_t = p$, set $z_{t+1} = \underline{p}$, and maintain this reduction in every period until inventory falls to a level less than or equal to c; at this time, production is increased back to p. (c) When X_t falls below or to b, $z_t = \bar{p}$, and this higher rate is maintained in every period until inventory moves to or above c; at this time production is decreased again to p. (d) The rate p is the minimum cost rate of operation. (e) The balance identity [equation (4)] holds.

More concisely:

$$(19)\quad z_{t+1} = \begin{cases} \underline{p} \begin{cases} \text{when } z_t > \underline{p} \text{ and } X_t \geq a \\ \text{" } z_t = \underline{p} \text{ " } X_t \geq c \end{cases} \\ p \begin{cases} \text{when } z_t = p \text{ and } a > X_t > b \\ \text{" } z_t = \bar{p} \text{ " } X_t > c \\ \text{" } z_t = \underline{p} \text{ " } X_t < c \end{cases} \\ \bar{p} \begin{cases} \text{when } z_t < \bar{p} \text{ and } X_t \leq b \\ \text{" } z_t = \bar{p} \text{ " } X_t \leq c; \end{cases} \end{cases}$$

$$(4)\quad X_t = X_{t-1} + z_t - \xi_t;$$

$$(20)\quad m(p) = \min [m(z_t)].$$

Six policy parameters determine production decisions; these are the production rates p, $\underline{p}$, and $\bar{p}$, and the inventory levels a, b, and c. The problem is to determine the set of values for these parameters which yield minimum cost of operation.

Two of the inventory levels, a and b, determine the range through which inventory fluctuates without signalling changes from the minimum cost rate of production. The level c is inserted between a and b so that after production has been changed away from p, the return to p will be made only when the system promises to remain in this minimum cost state for a reasonable time. If $\bar{p}$ is changed to p immediately after moving above b, there is a large chance of an immediate signal to switch back to

$\bar{p}$. The reasons for restricting the set of production alternatives to three elements will be investigated in the next chapter (Section III- 2.2.)

Once having decided upon a policy for controlling inventory and scheduling production, several important questions must be answered: can conditions be determined which suffice to assure the optimality of the policy? How does it compare in performance with other possible policies, under a variety of conditions? How are optimal decisions assured, subject to the condition that the policy is used? The last of these questions will be dealt with in the remainder of this chapter, and the other two will be considered in Chapter III.

Two analytic approaches, both mathematically rigorous, were described in Chapter I: these are dynamic programming and "steady state" analysis. Both deal successfully with policies having fewer decision rules than the (a,b,c) policy. As is seen in the Appendix, both computational and conceptual difficulties are encountered in the attempt to analyze (19-20) by these standard methods. It has been necessary to find approximations which describe the interactions between costs and policy parameters in the new policy; fortunately, properties have been discovered for stochastic processes similar to the one associated with the system (19-20). These properties are useful in obtaining approximation measures which make it possible to find optimal values for the policy parameters of the (a,b,c) policy.

3 A Stationary Analysis Based On Approximate Probability Measures

The difficulties which arise in using both standard analytic methods stem from the nonlinearity of the policy. The stationary inventory equations (55) comprise nine integral expressions: for each of the three

production states, it is necessary to know the probability of remaining in the present state, or moving to another state, since production is a factor which determines the transition probabilities. Fortunately, additional simplifying assumptions will enable the use of certain results, derived by Abraham Wald in his early work in sequential analysis, as surrogates for the information which otherwise would be obtained from the stationary inventory distribution. In the remainder of this chapter, some useful properties of random walks will be presented, and used to approximate the loss function of the model.[3]

3.1 Useful Results Regarding the Behavior of Cumulative Sums

The concept of a one-dimensional random walk provides an apt characterization of the operation of the (a,b,c) policy embedded in the model described in Section 2. In particular, when the inventory level is regarded as a cumulative sum, i.e.

$$X_n = \sum_{i=1}^{n} (z_i - \xi_i),$$

and the demands $\{\xi_i\}$ are independent, identically distributed random variables, the process is identical to one Wald has analyzed,[4] and his results may be applied directly.

Inventory changes between periods are given by

$$x_t = z_t - \xi_t.$$

Then, if $\Phi(\xi)$ is the distribution of the $\{\xi_i\}$, the distribution of the $\{x_i\}$ is obtained from the relation

---

[3] These approximations are conceptually similar to the analysis of Whitin and his predecessors. See the description of the rationale borrowed from them: supra, Section I-1.3.

[4] Abraham Wald: "On Cumulative Sums of Random Variables," Annals of Mathematical Statistics, Vol. XV, 1944, pp. 283-295.

(21) $P(x_i \leq u) = P(\xi_i \geq z_i - u)$

or $F_i(u) = 1 - \Phi(z_i - u)$. The distribution $F_i(x)$ is stationary provided $\Phi(\xi)$ is stationary and the $\{z_i\}$ are bounded; the first condition is assumed in the model, and the second is dictated by the (a,b,c) policy. Hence (21) may be written as

$$F(x) = 1 - \Phi(p - x)$$

to describe the distribution of inventory changes when the system is in the minimum-cost production state.

Wald's paper investigates two properties of random walks between two absorbing barriers a and b, where a or b is finite; these are the distribution of passage times, and the passage probabilities. The passage time is equal to n, where n satisfies the conditions

(22) $X_n \geq a$ or $X_n \leq b$, $a > X_j > b$ for $j = 0, 1, \ldots, n-1$.

In the context of the (a,b,c) inventory policy, it is the number of the time period in which inventory passes either of the decision levels a or b. The probabilities of a first passage at a and b, respectively are written P(a) and P(b), and satisfy the condition

(23) $P(a) + P(b) = 1$.

For notational convenience, two new variables are introduced:

$\alpha = a - X_0$

$\beta = X_0 - b$.

Let x_i be a random variable, with stationary density $\{x_i\}$. Let $X_0 = 0$; $\alpha > 0$ and $-\beta < 0$ are constants. Wald proves:

1. If Var (x) $\neq 0$, (22) must hold for some finite n, i.e. $P(n = \infty) = 0$.[5]
2. If E(x) and Var (x) both exist, and both are finite and nonzero; if there exists a δ such that

[5] ibid. lemma 1, p. 283

$P(e^x < 1 - \delta) > 0$

$P(e^x > 1 + \delta) > 0$;

if for any real h the moment generating function

$g(h) = E(e^{hx})$

exists, and the first two derivatives $g'(h)$ and $g''(h)$ exist: then, there exists a unique nonzero value of h, denoted H, which satisfies

(24) $E(e^{Hx}) = 1$.[6]

3. If D' is a subset of the complex plane such that $\psi(t) = E(e^{xt})$ exists and is finite for any element t of D', then the "fundamental identity"

(25) $E\{e^{X_n t}[\psi(t)]^{-n}\} = 1$

holds for all t in D'.[7] (It follows from the conditions imposed upon $\{x_i\}$ that this identity holds for all points t in the complex plane for which $\psi(t)$ exists and is greater than 1 in absolute value.)

4. If $\psi(t)$ exists for all real values of t, then (25) may be differentiated under the expectation sign any number of times with respect to t, at any value of t in the domain $|\psi(t)| \geq 1$.[8]

The fundamental identity and its derivatives lead directly to results which are useful in the present study.

The passage probabilities are obtained as follows.[9] Substitute H, as defined in (24), for t in the fundamental identity (25), to obtain

$E(e^{X_n H}) = 1$.

Let E_α be the value of this expectation under the condition $X_n \geq \alpha$, and E_β be the same expected value when $X_n \leq -\beta$. If $P(\alpha)$ is the probability $X_n \geq \alpha$, we have

(26) $P(\alpha)E_\alpha + [1 - P(\alpha)]E_\beta = 1$,

---

[6] ibid., lemma 2, p. 284 [7] ibid., p. 285

[8] This is the central theorem of A. Wald, "Differentiation Under the Expectation Sign in the Fundamental Identity of Sequential Analysis," Annals of Mathematical Statistics, Vol. XVIII, 1946, pp. 493-497.

or

(27) $P(\alpha) = (1 - E_\beta)/(E_\alpha - E_\beta)$.

When $|E(x)|$ and Var (x) are small compared to the range $\alpha + \beta$, a good approximation to (27) is

(28) $P(\alpha) \approx (1 - e^{-\beta H})/(e^{\alpha H} - e^{-\beta H})$, $E(x) \neq 0$.

The probability of a first passage at $-\beta$, $P(\beta) = 1 - P(\alpha)$, follows immediately from (26) and (28).

A condition imposed prior to the derivation of these results was $E(x) \neq 0$. It is seen, however, that as $E(x)$ tends to zero, H also tends to zero, and (28) converges, by l'Hospital's rule, to

(29) $P(\alpha) \approx \beta/(\alpha + \beta)$, $E(x) = 0$.

The expected passage times (durations before a change in production is signalled) are obtained by differentiating the fundamental identity with respect to t, at $t = 0$.[10] This differentiation yields

$$X_n - n[\psi'(0)]/\psi(0) = X_n - nE(x),$$

or

(30) $E(n) = E(X_n)/E(x)$, $E(x) \neq 0$.

In the sequel the notation $D(+)$, $D(0)$ and $D(-)$ will be used to represent the expected passage times when production is $\bar{p}$, p, and $\underline{p}$ respectively; these will also be referred to as durations in the plus, zero, and minus drifts. The approximation (28) may be used in evaluating (30):

(31) $E(n) \approx [\alpha(1 - e^{-\beta H}) - \beta(e^{\alpha H} - 1)]/(e^{\alpha H} - e^{-\beta H})E(x)$, $E(x) \neq 0$.

To evaluate $E(n)$ when $E(x) = 0$, (25) is differentiated twice with respect to t, at $t = 0$.[11] This second derivative is

(32) $\{[X_n - \frac{n\psi'(t)}{\psi(t)}]^2 - \frac{n\psi''(t)\psi(t) - [\psi'(t)]^2}{[\psi(t)]^2}\}\psi(t)^{-n} e^{X_n t}$.

---

[9] Wald, "On Cumulative Sums...," p. 287

[10] Wald, "Differentiation Under...," p. 494

[11] ibid., p. 496.

Since $\psi(0) = 1$, $\psi'(0) = E(x) = 0$, and $\psi''(0) = E(x^2)$, (32) becomes $X_n^2 - nE(x^2)$ at $t = 0$. This yields

(33) $E(n) = E(X_n^2)/E(x^2)$, $E(x) = 0$.

The approximation (28) is used to find

$$E(X_n^2) \approx \alpha^2 P(\alpha) + \beta^2 P(\beta),$$

yielding

(34) $D(0) \approx \alpha\beta/E(x^2)$.

The results (28) and (31) are not exact, since they use $P(X_n = \alpha)$ and $P(X_n = -\beta)$ as surrogates for $P(X_n \geq \alpha)$ and $P(X_n \leq -\beta)$; i.e. the "excess variable", the amount by which the variate X_n exceeds the barrier which it passes, is ignored.

3.2 An Approximation to the (a,b,c) Policy Objective Function

The approximate probability measures described in the preceding section provide the key to an analytically viable formulation of a loss function for the new model. An arbitrary planning horizon of T periods is established as a computational artifice; ignoring the initial state of the system seems less arbitrary when discussing the expected number of recurrences of an event over T periods, instead of the probability of an event in the coming period. Simple but appealing specific forms will be assigned to the four cost components of the loss function, and a procedure will be indicated whereby the loss function may be minimized.

The holding cost (17) is approximated by

(35) $$\int_0^\infty \eta(X)dF(X) \approx T\eta(a + b)/2,$$

where T is the planning horizon, η is the linear cost of holding one unit of inventory for one period, and $(a + b)/2$ is taken as an approximate measure of the expected value of the stationary inventory distribution.

Under the assumption that runouts never occur unless a passage below

b has first been detected, runout cost is formulated

$$(36)\quad \int_{-\infty}^{0} \rho(X)dF(X) \approx T\,\rho\left[\frac{e^{H(c-b)}-1}{e^{H(c-b)}-e^{Hb}}\right]P(b)/\gamma t$$

Here ρ is the "lump sum" cost incurred whenever inventory is found to be depleted at the end of a period; the bracketed expression is (30), the conditional probability of a passage below zero before a passage above c, given that production is $\bar{p}$.[12] The expression $P(b)/\gamma t$ is expected frequency of passage below b. The denominator, called the "cycle time" of the inventory process, is given by

$$\gamma t \approx D(0) + D(+)P(b) + D(-)P(a),$$

which is the sum of the expected durations in each drift, weighted by the probability of moving into that drift.[13] The numerator, P(b) is the portion of the cycle time spent in the plus drift.

The U-shaped ATUC curve (15) is represented by the parabolic form

$$(37)\quad m(z_t) = \frac{T\,m(\bar{p}-p)^2D(+)P(b) + m(p-\underline{p})^2D(-)P(a)}{\gamma t}$$

Here, the costs of production at $\bar{p}$ and $\underline{p}$ are weighted by the probability that the system is in the appropriate drift.

Production change costs include the costs of moving from the production rate p to either $\underline{p}$ or $\bar{p}$, plus the costs of subsequently

---

[12] The assumption that a runout cannot occur unless a passage of b has first been detected means that production is always in the state $\bar{p}$ at the time of a runout. When b is passed, production is changed to $\underline{p}$, and under our assumption, no runout occurs should the inventory level rise above c before it drops below zero, for then the system is restored to the zero drift. The bracketed term is the probability of a passage below zero while the system is still in the plus drift. In effect, equation (36) is the cost of a runout ρ, multiplied by the probability of a runout under the condition that a passage of b has occurred (the bracketed term) multiplied by the probability of a passage of b, which is $P(b)/\gamma t$.

[13] This formulation of the cycle time tacitly assumes that production never jumps from $\bar{p}$ to $\underline{p}$ or vice versa; every interval of production at $\bar{p}$ or $\underline{p}$ is preceded and followed by an interval of production at p. The cycle time, then, is simply the expected time required for inventory to move in the zero drift from c to a or b, and then move in the minus or plus drift back to c.

returning to p.[14] These costs are written in the form

$$(38)\quad T\left\{[\bar{K} + \bar{k}(\bar{p} - p)]\, P(b) + [\underline{K} + \underline{k}(p - \underline{p})]\, P(a)\right\} / \gamma t$$

where the terms in brackets are from (16), and $TP(b)/\gamma t$ and $TP(a)/\gamma t$ are the expected number of production increases and decreases over the planning period.

The expressions (35-38) constitute an explicit formulation of the costs assumed in the model, where all are written as functions of the appropriate parameters of the (a,b,c) policy. These four expressions are consolidated as follows:

$$(39)\quad L(a,b,c) = T\,\eta(a+b)/2 + T/\gamma t\Big\{[\,\rho(e^{H(c-b)}-1)/(e^{H(c-b)}-e^{-Hb})$$
$$+ \bar{K} + \bar{k}(\bar{p} - p) + m(\bar{p} - p)^2 D(+)]P(b)$$
$$+ [\underline{K} + \underline{k}(p - \underline{p}) + m(p - \underline{p})^2 D(-)]P(a)\Big\}$$

where η, ρ, $\bar{K}$, $\underline{K}$, $\bar{k}(\cdot)$, $\underline{k}(\cdot)$, $\bar{m}(\cdot)$, and $\underline{m}(\cdot)$ are cost parameters; a,b, and c are inventory decision parameters; and p, $\underline{p}$ and $\bar{p}$ are production decision variables. T is the planning horizon; and D(+), D(-), P(a), P(b) and γt are passage times and passage probabilities evaluated in Section 3.1. When these last expressions are written as explicit functions of the policy elements, (39) emerges as a transcendental equation which is too unwieldy to be optimized by the usual differentiation method. Hence, alternate procedures must be established for obtaining optimal values of decision variables.

3.3 Optimal Values for Two Policy Elements

In this section it is demonstrated that two of the decision deter-

[14] A single term may be used to represent the change in both directions because of the stationarity of the demand process; a change from p will be followed by a return to p with probability equal to one. As the length of the planning period T increases, the difference between the expected number of changes away from and returns to p is negligible.

minants, c and p, are easy to optimize, given certain nonrestrictive conditions.[15] These two policy elements must be selected so as to maximize D(0), the expected zero drift passage time, regardless of the explicit numerical values assigned to a and b. For suppose that, after selecting a and b, p and c were selected in such a way that D(0) was not maximized. Then, by finding values of p and c which increase D(0), it would be possible to reduce runout cost, production cost and production change cost all of which vary inversely with D(0), without increasing holding cost, which is independent of D(0). Only the values of p and c corresponding to maximum D(0) are not susceptible to such improvement.

Several additional conditions are imposed: (a) the demand distribution $\Phi(\xi)$ is normal with unit variance — $\Phi(\xi)$: $N(\mu', 1)$; (b) the approximation formulae (31) and (34) are taken as exact representations of random walk passage times; (c) the change costs encountered upon passing a are identical to those encountered upon passing b.[16] These conditions enable proof of the

---

[15] It was specified in defining the (a,b,c) policy that the production rate p should be the minimum cost rate (20). In this section, we will demonstrate further that p should be set equal to the mean of the demand distribution, $E(\xi)$. Perhaps it seems that it will be necessary to perform some sleight-of-hand to assure $\min [m(z)] = E(\xi)$. We can, of course, simply assume that this equality exists; however the assumption is stronger than necessary, for the assumption of demand stationarity, in combination with entreprenurial rationality strongly suggests that (a) if a firm has been in operation over a long period of time, it will have adjusted its capital equipment acquistions in a manner that conforms to the objective of filling expected demands at minimum cost; (b) if a firm has just begun operation, the assumption that the demand distribution is known immediately enables satisfaction of the equality. An alternative (but again, unnecessarily strong) assumption is that the demand function (in which expected demand is plotted against price) is of unit elasticity throughout its length; the desired equality is attained through adjustment of the price level.

[16] It is believed that the theorem holds for a wide variety of continuous distribution functions, although only the normal and exponential ($F(x) = \int_0^\infty xe^{-x}$) have been tried. The assumption of unit variance is convenient computationally. Assumption (c) is discussed at length later (infra. Section III-3).

Theorem. Optimal values of c and p in the (a,b,c) inventory policy are respectively $(a+b)/2$, the midpoint of the zero-drift inventory range, and μ^*, the mean of demands.

Proof. The notation of Section 3.1, specifically $\alpha = a-c$ and $\beta = c-b$, and the balance relation (4), are recalled. It follows from (4) that the inventory changes $\{x_i\}$ have the distribution $F(x) : N(\mu,1)$, given the assumption about the normality and stationarity of demands, where

(40) $\mu = p - \mu^*$.

It is necessary to demonstrate that no selection of c is more advantageous than $(a+b)/2$, and no selection of p is more advantageous than μ^*. Let $2r = \alpha+\beta$ represent the distance between the upper and lower inventory control levels. Having shown the necessity to maximize D(0) for fixed a and b, it will suffice to prove the inequality (41) always holds. Substitute $2r-\beta$ for α in (31), and let $\alpha = \beta$ in (34) to obtain

$$(41) \qquad \frac{(2r-\beta)(1-e^{-\beta H}) - \beta(e^{(2r-\beta)H}-1)}{\mu(e^{(2r-\beta)H} - e^{-\beta H})} \leq r^2$$

where the left side represents the expected first passage time for arbitrary α, β and μ, while the right side represents the expected first passage time for $\alpha = \beta = r$, and $\mu = 0$. H was defined in (24).

The moment generating function of a normally distributed variate is

$$(42) \qquad E(e^{hx}) = e^{(h\mu + \frac{1}{2}\sigma^2 h^2)}$$

where $\sigma^2 = \mathrm{Var}(x)$; setting (42) equal to 1 yields, for $\sigma^2 = 1$,

$$(43) \qquad \mu H + (1/2)H^2 = 0, \text{ or } H = -2\mu.$$

Now (41) may be rewritten

$$(44) \qquad (2r-\beta)(1-e^{2\beta\mu}) - \beta(e^{-2(2r-\beta)\mu} - 1) \geq r^2\mu(e^{-2(2r-\beta)\mu} - e^{2\beta\mu}).$$

The inequality is reversed, since the denominator of (41) is negative, regardless of the sign of μ. To determine whether the inequality (44) indeed holds, the right term is transposed, yielding

(45) $(2r-\beta)(1-e^{2\beta\mu}) - \beta(e^{-2(2r-\beta)\mu}-1) - r^2\mu(e^{-2(2r-\beta)\mu} - e^{2\beta\mu})$

on the left side. If the inequality (44) is to hold, (45) must take a nonnegative maximum value. Differentiation with respect to μ yields

(46) $e^{2\beta\mu}[2\mu\beta r^2 + r^2 - 2\beta(2r-\beta)] - e^{-2(2r-\beta)\mu}[r^2 - 2\mu r^2(2r-\beta) - 2\beta(2r-\beta)] = 0.$

We see that (46) has a root $\mu = 0$. This root is unique: rewriting the derivative as

$$e^{4r\mu} = \frac{r^2 - 2\beta(2r-\beta) - 2\mu r^2(2r-\beta)}{r^2 + 2\mu\beta r^2 - 2\beta(2r-\beta)}$$

shows that when $\mu > 0$, the left side > 1 and the right side < 1. The reverse inequalities hold for $\mu < 0$. To establish sufficiency, take Maclaurin expansions of the exponential terms of (44), obtaining

(47) $\mu^2[-2\beta\mu(2r-\beta) - 2\beta(2r-\beta)^2 + 2r^2(2r-\beta) + 2\beta r^2] - \mu^3[-2r^2(2r-\beta)^2 + 2r^2\beta^2].$

If (47) takes a minimum value at $\mu = 0$, this expression must be positive everywhere in the neighborhood of the root. Since μ may be taken arbitrarily small, the effect of terms of successively higher powers in μ becomes unimportant in determining the sign of (47). Focusing on terms in μ^2, it is seen that they are not dependent upon the sign of μ. The expression within the first bracket of (47) may be rewritten $-8\beta^2 r + 4\beta^3 + 4r^3$, which is always nonnegative for positive β and r. Thus (47) takes its minimum at $\mu = 0$, $\beta = r$, verifying inequalities (44) and (41) and completing the proof.

3.4 The Objective Function Reexamined

The results of the last section reduce the optimization problem to one of finding values for four decision parameters, instead of six, and in addition, they disperse much of the difficulty encountered in determining the remaining policy elements. This is seen in the changes which appear in the

loss function.

The passage times in the plus and minus drifts are obtained from (31). Some interpretation is called for: in the plus drift, the mean change in the inventory level between periods is given by $\bar{\mu} = \bar{p} - \mu^*$, $\bar{\mu} > 0$. In this drift, the α of (31) represents the distance (c - b); for it is assumed that the process begins its walk in the new drift at the lower barrier b, and the property of interest is the time required to move back to the center of the (a - b) range. Further, since there is no lower barrier of importance (passages below 0 are of interest only in dealing with runout costs, and not in dealing with expected duration in the plus drift), the β of (31), representing the distance the walk may proceed in the negative direction without encountering a barrier, approaches infinity. On this basis, it is seen

$$D(+) = \lim_{\beta\to\infty} \left[\frac{\alpha(1-e^{2\bar{\mu}\beta}) - \beta(e^{-2\bar{\mu}\alpha}-1)}{\bar{\mu}(e^{-2\bar{\mu}\alpha} - e^{2\bar{\mu}\beta})} \right] = \alpha/\bar{\mu} = r/\bar{\mu}.$$

Similarly, $D(-) = \beta/\underline{\mu} = r/\underline{\mu}$, where $\underline{\mu} = \mu^* - \underline{p}$. These results, with (28), (29), (34) and (42) make it possible to write (39) as an explicit function of the decision parameters:

$$(48) \quad L(a,b,c) = T\left\{\eta(r+b) + \frac{(\rho/2r)\,\dfrac{(e^{-2\bar{\mu}r}-1)}{(e^{-2\bar{\mu}r} - e^{2\bar{\mu}b})} + K/r + \tfrac{1}{2}(k/r + m)(\bar{\mu}+\underline{\mu})}{r + 1/2\bar{\mu} + 1/2\underline{\mu}}\right\}$$

where 2r = a-b, the width of the zero drift inventory range in standard deviations of demand.

Before the question of obtaining optimal values of decision variables from the expression (48) is considered, it must be determined whether (48) is a suitable form for the loss function.

3.5 Modifications in Light of Limiting Behavior

Two sorts of difficulty may arise which jeopardize the results obtained from equation (48): (a) The emprical cost parameters might posess some

pathological structure which forces one or more of the decision variables to take infinite values. For example, if the runout cost parameter ρ is infinitely large, or the holding cost parameter η is zero, the upper inventory barrier of the (a,b,c) policy would never stabilize at a finite value. (b) The approximation forms assigned to these costs might misrepresent their behavior, i.e. certain of the cost components might not respond to changes in one of the decision variables in the way it should: holding cost not increasing with increases in r, for example. The contingency (a) may be precluded by assumption — the (a,b,c) policy simple is not used for dealing with such cost structures. Contingency (b), which is considered in this section, is seen to arise when total cost accruing from use of the (a,b,c) policy is represented by (48), and thus some modification is necessary.

As a preliminary, it is established that the loss function (48) is everywhere continuous when the four decision variables are restricted to positive values. This desirable condition obtains because the function is composed only of sums and products of linear and exponential terms; further, no denominator in the function can vanish.

The function is composed of four elements: the term $\eta(r+b)$ is holding cost per period; the exponential expression in the numerator is the approximation to runout cost per period, first presented in equation (36); the terms $K/r + k/2r(\bar{\mu} + \underline{\mu})$ in the numerator represent production change costs (38), and the term $(m/2)(\bar{\mu} + \underline{\mu})$ represents production cost (37).

Two troublesome concomitants of the representation (48) are discovered; holding r,b, and $\underline{\mu}$ constant it is seen that

$$(49)\qquad \lim_{\bar{\mu}\to 0} L(a,b,c) = \eta(r+b) + \frac{\rho/2(r+b) + K/r + \frac{1}{2}(k/r + m)\underline{\mu}}{r + 1/\bar{\mu} + 1/\underline{\mu}} \to \eta(r+b)$$

unfortunately, the function $L(a,b,c)$ takes its minimum when $\bar{\mu} = 0$, since this is the only value which causes the cost of runout and production associated costs to vanish. This representation conflicts with intuition regarding the behavior of cost in the system, for unless $\bar{\mu} > 0$, the probability that the system is perpetually out of stock approaches one, meaning that substantial runout costs are ignored in equation (49).

The second difficulty is perceived upon considering

$$\lim_{\underline{\mu} \to 0} \quad L(a,b,c) \to \eta(r+b)$$

with $\bar{\mu}$, r, and b held constant. Again a minimum value occurs when the decision variable is set equal to zero. But if $\underline{\mu}$ is set equal to zero, the probability of passing any arbitrarily high inventory level becomes 1, and holding cost explodes.

In order to ameliorate these weaknesses, the loss function (48) is rewritten

$$(50) \quad L(a,b,c) = \left[\Theta(r+b)/\underline{\mu} + \frac{(\upsilon/\bar{\mu})(e^{-2\bar{\mu}r}-1)/(e^{-2\bar{\mu}r}-e^{2\bar{\mu}b}) + K + \frac{1}{2}(k+mr)(\bar{\mu}+\underline{\mu})}{r(r+1/2\bar{\mu}+1/2\underline{\mu})}\right]T$$

where

$$\Theta = \begin{cases} \eta O(\underline{\mu}) & \underline{\mu} \to \infty \\ \eta O(1/\underline{\mu}) & \underline{\mu} \to 0 \\ \eta\underline{\mu} & \text{otherwise} \end{cases}$$

$$\upsilon = \begin{cases} \rho O(\bar{\mu}) & \bar{\mu} \to \infty \\ \rho O(1/\bar{\mu}) & \bar{\mu} \to 0 \\ \rho\bar{\mu} & \text{otherwise;} \end{cases}$$

where $O(\mu) \to$ a constant as $\mu \to \infty$.

These forms assure that cost of runout and of holding react appropriately to any values taken by the production change $\bar{\mu}$ and $\underline{\mu}$.

It is easy to verify that after these modifications, the loss function in the form (50) is still continuous.

The function (50) behaves properly as the decision variables individually approach their limiting values. In the following tests of limiting behavior, each decision variable is permitted to approach its limiting values, while the other three variables are held constant at some finite nonzero level. At the following six limits, the function (50) is well-behaved:

$$\lim_{r\to\infty} \quad L(a,b,c) \approx \Theta(r+b)/\bar{\mu} \quad \to \infty$$

$$\lim_{r\to 0} \quad L(a,b,c) \approx \frac{K + \frac{1}{2}k(\bar{\mu}+\underline{\mu})}{r(r+1/2\bar{\mu}+1/2\underline{\mu})} \quad \to \infty$$

$$\lim_{\bar{\mu}\to\infty} \quad L(a,b,c) \approx \frac{\frac{1}{2}(k+mr)(\bar{\mu}+\underline{\mu})}{r(r+1/2\underline{\mu})} \quad \to \infty$$

$$\lim_{\underline{\mu}\to 0} \quad L(a,b,c) \approx \Theta(r+b)/\underline{\mu} \quad \to \infty$$

$$\lim_{\underline{\mu}\to\infty} \quad L(a,b,c) \approx \frac{\frac{1}{2}(k+mr)(\bar{\mu}+\underline{\mu})}{r(r+1/2\bar{\mu})} \quad \to \infty$$

$$\lim_{b\to\infty} \quad L(a,b,c) \approx \Theta(r+b)/\underline{\mu} \quad \to \infty$$

In considering

$$(51) \quad \lim_{b\to 0} \quad L(a,b,c) = \Theta\, r/\underline{\mu} + \upsilon/\bar{\mu} + \frac{(\text{production cost})}{r(r+1/2\bar{\mu}+1/2\underline{\mu})}$$

it is necessary to determine whether an optimal policy could reasonably specify a value of $b > 0$; hence a comparison is made between (51) and

$$[L(a,b,c)\,|\,b=1] = \Theta(r+1)/\underline{\mu} + \frac{\left[\frac{(\upsilon/\bar{\mu})(e^{-2\bar{\mu}r}-1)}{(e^{-2\bar{\mu}r}-e^{2\bar{\mu}})}\right] + (\text{production cost})}{r(r+1/2\underline{\mu}+1/2\bar{\mu})}$$

If the loss function is to take a smaller value when $b = 1$ than when $b = 0$, it is necessary that

$$\Theta/\underline{\mu} < [1 - (e^{-2\bar{\mu}r}-1)/(e^{-2\bar{\mu}r}-e^{2\bar{\mu}})](\upsilon/\bar{\mu})/r(r+1/2\underline{\mu}+1/2\bar{\mu})$$

which is reasonable condition. For instance, let $\underline{\mu} = r = 1$, $\bar{\mu} = \frac{1}{2}$, to

obtain $\Theta < (3/5)v$; the cost of a runout must be more than five thirds as great as the cost of holding one unit of stock for one period.

Similarly, when

$$\lim_{\bar{\mu}\to 0} L(a,b,c) = (\text{holding cost}) + 2v/(r + b)$$

and

$$[L(a,b,c)\,|\,\bar{\mu} = 1] = (\text{holding cost}) + \frac{v(e^{-2r}-1)/(e^{-2r}-e^{2b}) + K + \frac{1}{2}(k+mr)(\underline{\mu}+1)}{r(r + 1/2\underline{\mu} + 1/2\bar{\mu})} \Big/ r(r+1/2\underline{\mu}+1/2\bar{\mu})$$

For $L(a,b,c)$ to be less when $\bar{\mu} = 1$ than when $\bar{\mu} = 0$, it is necessary that

$$\frac{K + \frac{1}{2}(k + mr)(\underline{\mu} + 1)}{r(r + 1/2\underline{\mu} + 1/2)} < v[2/(r+b) - (e^{-2r}-1)/(e^{-2r}-e^{2b})r(r+1/2\underline{\mu}+1/2)].$$

Again, this is a reasonable condition; for instance, let $r = \underline{\mu} = b = 1$ to obtain $(K + k + m)/2 < 6v/7$.

Expression (50) is a more suitable approximation to the loss function of the (a,b,c) policy than is expression (48), because the former conforms more closely to the way in which costs are expected to behave as the decision variables move toward the limits of their ranges.

The usual analytic procedure for finding simultaneous minimizing values of four decision variables is cumbersome, even when the function involved is simple in form. When confronted by so untidy a form as (50), this optimization process is nearly useless. However, the minimization problem is solvable by numerical methods on a high-speed computer;[17] (50) is similar to many seemingly intractable "engineering functions" which have been optimized with the aid of computers.

---

[17] A code has been developed for the IBM-704 whereby expressions as complex as (50) can be optimized. This program, which is designed for multivariate regression analysis, is described at length in G.E.P. Box, The Use of Statistical Methods in the Elucidation of Basic Mechanisms, Princeton University, Statistical Techniques Research Group, Technical Report no. 7, October, 1957. I am indebted to Roger Pinkham for this reference.

That the policy parameters obtained by maximizing this loss function are not unreasonable is verified by a small numerical example. No attempt will be made to locate optima, for while the gradient methods whereby this may be accomplished are easy to use when a computer is available, they do not lend themselves readily to hand computation. Rather, it will be demonstrated that the loss function is quite "flat" over a surprisingly large range of combinations of decision variables, which encourages the belief that the policy may be useful even when a computer is not available.

Suppose the cost parameters are: $\eta = 1$, $\rho = 1000$, $K = 300$, $k = 10$, and $m = 5$. The parameters ρ and K are fixed charges, the others are costs per unit, where the unit is equal to the standard deviation of the demand distribution; e.g. it costs 1 to hold σ_ξ items of finished goods for one period, and 10 to make a change of σ_ξ units per period in the production rate. Then, when the decision variables $(r,b,\bar{\mu},\underline{\mu}) = (1,1,1/4,1/4)$, the total cost is approximately $2 + \frac{1000(.3935/1.0422) + 300 + 15/4}{2}$, or about 342.66.

The following table shows the approximate total cost associated with several possible combinations of $(r,b,\bar{\mu},\underline{\mu})$:

(3,2,1/2,1/2)	34.46	(8,5,1/2,1/2)	17.15
(5,3,1,1)	19.25	(10,6,1/2,1/2)	18.77
(5,3,1/2,1/2)	18.48	(10,6,1,1)	19.27

Since no combination involving $r + b > 18$ can be optimal, the best choice will be somewhere in the range covered in the table. Notice that, although all decision variables have been subjected to 100 per cent variations in magnitude, the cost variation between the worst and the best of the last five combinations is less than 12.5 per cent.

With the objective function in a form which is clearly amenable to optimization, attention is shifted to the topic of production costs and production change costs in Chapter III.

APPENDIX[1]

FORMULATIONS BASED UPON STANDARD ANALYTIC APPROACHES

1 Dynamic Programming Formulation of the (a,b,c) Policy

The expected one-period costs of operation when the (a,b,c) policy is used depend upon the initial production state z_1 and upon the initial inventory X_0. These costs are given by

$$
(52)\quad \begin{cases}
R(X_0,p) = \eta[X_0 + p - \int_{-\infty}^{X_0+p} \xi d\Phi(\xi)] + \rho\int_{X_0+p}^{\infty} d\Phi(\xi) + [K + k(p-\underline{p})]\int_{-\infty}^{X_0+p-a} d\Phi(\xi) \\
\qquad\qquad + [K + k(\bar{p}-p)]\int_{X_0+p-b}^{\infty} d\Phi(\xi) . \\
R(X_0,\bar{p}) = \eta[X_0 + \bar{p} - \int_{-\infty}^{X_0+\bar{p}} \xi d\Phi(\xi)] + \rho\int_{X_0+\bar{p}}^{\infty} d\Phi(\xi) + m(\bar{p}-p)^2 \\
\qquad + [K + k(\bar{p}-p)]\int_{X_0+\bar{p}-a}^{X_0+\bar{p}-c} d\Phi(\xi) + [K + k(\bar{p}-\underline{p})]\int_{-\infty}^{X_0+\bar{p}-a} d\Phi(\xi) . \\
R(X_0,\underline{p}) = \eta[X_0 + \underline{p} - \int_{-\infty}^{X_0+\underline{p}} \xi d\Phi(\xi)] + \rho\int_{X_0+\underline{p}}^{\infty} d\Phi(\xi) + m(p-\underline{p})^2 \\
\qquad + [K + k(p-\underline{p})]\int_{X_0+\underline{p}-c}^{X_0+p-b} d\Phi(\xi) + [K + k(\bar{p}-\underline{p})]\int_{X_0+\underline{p}-b}^{\infty} d\Phi(\xi) .
\end{cases}
$$

In each of these equations, the first term is holding cost (17) formulated as a linear function of ending inventory; the second term is runout cost (18), formulated as a lump sum penalty incurred when ending inventory $X_1 < 0$; the third term is production cost (15), formulated as a quadratic function of the deviation from the minimum-cost level; and the remaining terms are the production change costs (16).

In addition to current period costs, the loss function must evaluate

[1] The reader is advised to skip to Chapter III unless he is specifically interested in these analytic approaches.

the impact of a decision in the current period upon future operations. By the principle of optimality, the discounted future cost stream may be summarized as follows:

$$(53)\quad L(X_0, \{p\}) = R(X_0, \{p\}) + \alpha E[L(X_1, \{p\})], \quad \{p\} \in (\bar{p}, p, \underline{p}).$$

Upon evaluating the second term on the right-hand side, the total effect of (a,b,c) policies upon expected costs will be known. This evaluation is accomplished by weighting total future costs for each production state by the probability of being in that state:

$$(54)\begin{cases} E[L(X_1,p)] = L(X_1,\underline{p})\int_{-\infty}^{X_0+p-a} d\Phi(\xi) + L(X_1,p)\int_{X_0+p-a}^{X_0+p-b} d\Phi(\xi) + L(X_1,\bar{p})\int_{X_0+p-b}^{\infty} d\Phi(\xi) \\ E[L(X_1,\bar{p})] = L(X_1,\underline{p})\int_{-\infty}^{X_0+\bar{p}-a} d\Phi(\xi) + L(X_1,p)\int_{X_0+\bar{p}-a}^{X_0+\bar{p}-c} d\Phi(\xi) + L(X_1,\bar{p})\int_{X_0+\bar{p}-c}^{\infty} d\Phi(\xi) \\ E[L(X_1,\underline{p})] = L(X_1,\underline{p})\int_{-\infty}^{X_0+\underline{p}-c} d\Phi(\xi) + L(X_1,p)\int_{X_0+\underline{p}-c}^{X_0+\underline{p}-b} d\Phi(\xi) + L(X_1,\bar{p})\int_{X_0+\underline{p}-b}^{\infty} d\Phi(\xi) \end{cases}$$

These expressions are explained as follows: if production is in the state p, and sales are less than $X_0 + p - a$, we move into the state $\underline{p}$; after this move, the future expected costs become $L(X_1,\underline{p})$. By combining (52) and (54), three equations of the form (53) are obtained. The objective in dealing with such a system is to find values of the decision variables which yield minimum cost. The corresponding (but simpler) set of relations for the s,S policy is solved by Arrow, Harris and Marschak; they use boundary conditions to transform their equations into a renewal equation.[2] Thus far the author has been unable to solve the recursive functionals (52) of the (a,b,c) policy.

2 The Stationary Inventory Distribution of The (a,b,c) Policy

The policy rules of Section 2.1, combined with the stationary, random

[2] "Optimal Inventory Policy" op.cit. The functional equations (52) differ from the general dynamic programming formulation (8) since they contain no max operator. This is because the form of the policy is prespecified: the decision is completely determined when the policy parameters are established, and the variables with respect to which we maximize are known in advance.

demand distribution assumed in the model, determine a Markovian inventory process. The stationary inventory distribution[3] comprises three parts:

$$F(X) = F^{-}(X) + F^{o}(X) + F^{+}(X),$$

where the successive components describe stationary behavior under the condition that the system is in the minus, zero, and plus drift respectively. The component terms are written:

$$dF^{-}(X) = \int_{c}^{\infty} \varphi(X' + \underline{p} - X)dF^{-}(X') + \int_{a}^{\infty} \varphi(X' + \underline{p} - X)dF^{o}(X') + \int_{a}^{\infty} \varphi(X' + \underline{p} - X)dF^{+}(X')$$

$$(55)\quad dF^{o}(X) = \int_{b}^{c} \varphi(X' + p - X)dF^{-}(X') + \int_{b}^{a} \varphi(X' + p - X)dF^{o}(X') + \int_{c}^{a} \varphi(X' + p - X)dF^{+}(X')$$

$$dF^{+}(X) = \int_{-\infty}^{b} \varphi(X' + \bar{p} - X)dF^{-}(X') + \int_{-\infty}^{b} \varphi(X' + \bar{p} - X)dF^{o}(X') + \int_{-\infty}^{c} \varphi(X' + \bar{p} - X)dF^{+}(X')$$

The first term in the first equation states that if X', beginning inventory n periods in the future (where n is a large number) is greater than c, and the production $n-1$ periods in the future is $\underline{p}$, then production in the period n will be $\underline{p}$, and the probability of having exactly X units on hand at the end of the period is $\varphi(X' + \underline{p} - X)$, where $\varphi(\xi)$ is the demand density. Thus the first equation is the density function of ending inventory when production during the period was $\underline{p}$. Distribution functions are obtained by integrating each equation over the range of $\varphi(\xi)$; taking this step, we see terms will be of the form

$$F^{-}(X) = \int_{c}^{\infty} \left[\int_{-\infty}^{\infty} \varphi(X' + \underline{p} - X)d\xi \right] dF^{-}(X') + \text{etc.}$$

These equations have thus far resisted solution, even when simple forms of the density $\varphi(\xi)$ are assumed, as was discovered when $\varphi(\xi) = e^{-\xi}$ and $\varphi(\xi) = \xi e^{-\xi}$ were tried. Certain conditions must be satisfied, of course, i.e $F(X) = 1$, $F^{-}(X) = F^{o}(X + p - \underline{p})$, etc. but these add little to our ability to solve the equations.

[3] A standard method exists for determining this distribution when the inventory process is Markovian: cf. Arrow, Karlin and Scarf, op. cit. p. 225.

The optimization problem is not completely solved when the stationary inventory distribution is determined, for unlike the s,S policy[4] not all of the costs of the (a,b,c) policy can be measured in terms of the stationary inventory distribution. The cost of changing the production rate does not lend itself to measurement in terms of this stationary distribution, and some alternative method must be obtained to describe the expected frequency of different types of change, such as the one described in Section II-3.1.

---

[4] The costs of the s,S policy were related to the stationary inventory distribution which it determines: Supra. Section I-1.3.

CHAPTER III

DYNAMIC AND COMPARATIVE STATIC COSTS OF PRODUCTION

1 The Production Cost Function and the Employment Level

In this chapter the factors which underlie the specific forms of production cost and production change cost that appear in the model of Chapter II are explored [see equations (15) and (16)]. To be able to cope with production change cost as it is there formulated is the principal reason for adopting the (a,b,c) policy, so the reasons which underlie the formulation of this cost constitute a significant portion of the rationale of the policy. Static production costs, similar to those described in the economic literature on production,[1] also play a substantial role in shaping the policy: they dictate that production rates be kept within a certain range, and make it desirable to produce at the minimum cost rate.

Two conditions assumed in the model (Section II-2) are important to the formulation of production cost: the assumption that demand is stationary, and the assumption that the decisionmaker does not control factors affecting demand, such as price. This last assumption implies that cost minimization is a valid objective; the first means that the firm will never find it necessary to move to another short run cost curve, provided

[1] Cf. Sune Carlson, A Study in the Pure Theory of Production (London: P.S. King and Son, 1939). This theory proceeds from the assumption that the lowest cost input mix for any desired rate of output has been determined; Chenery has studied the problem of determining this optimal input mix. Cf. Hollis Chenery, "Engineering Production Functions," Quarterly Journal of Economics, Vol. LXIII, 1949, pp. 507-531.

the present capital structure is on the average capable of meeting customer demands. These two assumptions enable the problem of inventory-production programming to be divorced completely from related problems of capital planning.

This study emphasizes short run considerations which are analagous to the long run problem of capital planning, namely the planning of employment levels. Comparisons are drawn between "intermediate run" and short run alternatives: direct labor will first be regarded as fixed (the short run case), then variable (intermediate run), and the implications of the two possibilities will be explored. A production change in which production workers are hired or fired is a move to a new short run curve, along the intermediate run curve; while a change in which the direct labor force is unchanged is a short run adjustment. The intermediate run function as defined above is a short run function as defined traditionally, and the term 'short run' has been given a more restrictive definition. These intermediate run and short run average cost schedules show the minimum costs of operation at each possible rate of production, when capital equipment is fixed, and given the alternative of varying or not varying the labor force.

Though there will doubtless be sharp differences in the configuration of short run and intermediate run costs in different situations, the following observations may be valid under widely varied circumstances. To increase output without hiring, one or another of the following alternatives may be available in a given situation: (a) overtime production; (b) reduction of maintenance, cleanup or other necessary parts of the operating routine which may be neglected temporarily; (c) increased operating rate, with no change in the labor hour input. Diseconomies encountered

in pursuing this last alternative may be attributed to employee fatigue, deterioration of equipment, increased cost of spoilage or scrap, or reduced quality. A short run reduction in output leads to costs of "undertime", which occurs when the work force is diverted into nonproductive pursuits, e.g. wall and window washing.

Intermediate run adjustments create entirely different problems. By varying the labor force, lower direct labor cost per unit is obtained than when the labor force is unchanged. In the intermediate run, increases in production larger than some critical size will probably be effected by an additional shift, whenever possible. Changes smaller than the critical size may be effected by re-balancing crews on existing shifts, if greater labor intensity can lead to an increase in output; or all nonproductive routines, such as maintenance, cleanup, etc. may be performed after regular hours; this plan may yield substantially more production at the expense of hiring a cleanup and maintenance crew to come in after hours. Each of these methods of increasing production will result in higher direct labor cost per unit, if prior to the change, the firm was operating at the minimum cost level for both the short and the intermediate run.[2] If a firm is operating more than one shift, an intermediate run reduction may be effected by the layoff of an entire shift; or by laying off some workers, reassigning those remaining, and continuing operation of the same number of shifts as before the decrease. Again, unit costs will be higher at the new operating level than at the equilibrium level. When only one shift is being operated, only the layoff and reassignment recourse remains open.

The same reasoning which has led economists to draw the long run ATUC

[2] This is the condition which we have assumed to exist in the preceding chapter; supra. p.33, note 15.

curve as the envelope to a family of short run ATUC curves, underlies the formulation of the intermediate run curve as the envelope to a family of (redefined) short run curves: more efficient input mixes can be realized when all the factors are variable. The U-shape assumed for the production cost element (15) is based upon this adaptation of the arguments of the economic theory of production.

2 Costs of Changing Production Rates

If only comparative static cost considerations entered into production and employment decisions, employment would doubtless oscillate more than it does. However, when output is changed, the path of adjustment from the initial to the final equilibrium point is not along the comparative static production cost function $m(z)$, regardless of whether an intermediate run or short run adjustment is called for.[3]

This accepted doctrine regarding dynamic cost behavior is reflected in the production change costs formulated in our model (16). A great deal has been learned about these costs by industrial engineers, in their experience with such widely-practiced devices as cost control, time and motion study, methods analysis, and learning studies. These programs all deal directly with psychological and physiological factors which underlie the dynamic costs of changing rates.

The factors underlying these costs divide naturally into three classes: (a) time consumed in calculating the minimum cost input combination at the

[3] This was pointed out succinctly by W.W. Cooper: "Cost curves ... cannot be specified without reference to the time allowed for adjustment. For diagrammatic convenience ... this fact is often ignored, but it must be carefully taken into account for purposes of empirical investigation. In proceeding from one output level to another, the entrepreneur (in the theory of comparative statics) can move smoothly along the 'optimum' curve only when ample time is allowed for adjustment. Otherwise the firm will experience higher costs, even though at the new output level and with more

new production rate, which we call input engineering costs (b) time required for labor to find and to grow accustomed to the most efficient patterns of motion at the new output rate, the so-called "learning costs" (c) costs of paperwork and managerial effort in reassigning and perhaps altering the size of the work force, calculating optimal input combinations and in general preparing the production system for the proposed change, the administrative costs.

Chenery describes some of the difficulties encountered in reducing the first type of cost.[4] These include (a) problems of description: identifying all possible technological combinations capable of yielding the specified output rate (b) input price changes, which upset previously completed "optimal input combination" calculations, and (c) technological changes, which result in new processes or materials, and alter the profile of the set of feasible input possibilities. In the face of the large number of possible output rates, the large number of input combinations associated with each output rate, and the frequency of changes in input cost and technology, the maintenance of accurate information regarding the current status of the ATUC function would be very nearly impossible. When a change to a new output rate is called for, probably the most economical and practical method of making the change is to estimate the economical input combination,

---

[3](cont.) time subsequently at its disposal, it may, under appropriate conditions, effect the necessary adjustments and reduce costs to the optimum curve. In short, there is not just one cost curve but rather a family of curves, each appropriate to a particular time period and rate of adjustment at a new output level." Cf. "Extending the Theory of the Firm," Quarterly Journal of Economics, Vol. LXV, 1951 pp.87-109.

[4] Op. cit. The separation of labor from other inputs in considering the problems of production dynamics is suggested by this remark of Chenery's: "Engineering as commonly conceived, concerns itself principally with the performance of machines rather than of men. Consequently, the accuracy of a production function based on engineering data alone will vary inversely with the variability of the labor input." (p. 510).

and then to design and perform experiments which indicate cost-reducing variations in the estimated input combination.[5]

The second factor to which a causal role has been assigned in the emergence of production change costs, is the time until the labor force begins operating in the most efficient possible fashion. Industrial engineers have long been concerned with the identification of optimal motion patterns for the individual worker in the production system.[6] Such a pattern will be repetitive, and maintained at a constant rate and rhythm. If a group of workers carries out the same set of tasks at the same speed every day, their performance level should approach an "ideal" which is determined by time study. Conversely, if crew assignments are switched, rates of operation are varied, and individual rhythms are upset, the ideal performance is an unfair standard.[7]

Psychologists have recently begun to devote attention on a large scale to mathematical models of learning processes.[8] To the extent that these models are reliable representations of actual learning processes, they provide reassuring support to our generalizations that time is

---

[5] The success of such experiments will depend upon the smoothness and convexity of the cost surfaces, otherwise local, rather than global, optimum input combinations may be discovered.

[6] For a description of this topic, see Marvin Mundel, "Motion and Time Study" Section 5 of W.G. Ireson and E. L. Grant (editors), Handbook of Industrial Engineering and Management (Englewood Cliffs, N.J.: Prentice-Hall, 1955).

[7] In highly repetive processes, fatigue may be an important consideration. In multi-worker operations, where there are numerous diverse jobs requiring the same level of skill, the workers may rotate jobs frequently in order to avoid fatigue.

[8] Cf. Robert R. Bush and Frederick Mosteller, Stochastic Models for Learning (New York: John Wiley and Sons, 1955), or R.R. Bush and William K. Estes (editors), Studies in Mathematical Learning Theory (Stanford, California: Stanford University Press, 1959).

required for workers to assimilate the different patterns of operation when the output rate is changed. By assigning appropriate specific meanings to the dependent and independent variables of a model of Bush and Mosteller,[9] the impact of learning time upon production change cost becomes clear. Let the stimulus be the specification of output volume in the coming production scheduling period, the response be the selection by the worker of his entire sequence of motor actions during the period, and the environmental event be the effectiveness of the employee's performance. The approach to optimal effectiveness is determined by a probability rule, such as

> the probability that optimal effectiveness is realized is proportional to the number of times the identical stimulus (production rate specification) was received in the past, and inversely proportional to the number of periods since the identical stimulus was last received. 10

This sort of rule is consistent with other reformulations of the basic Bush-Mosteller model cited in footnote 9; in most of these the response probability increases with the frequency of stimulus. The simple model proposed here is designed to provide rationale for the policy rules (19), and simultaneously to be consistent with the formulation of production change costs (16). That this can be done so readily by use of a simple mechanism which may possess substantial descriptive validity is reassuring.

Administrative costs of production change are due to these necessary

---

[9] Bush and Mosteller, "A Mathematical Model for Simple Learning," Psychological Review, Vol. 58, 1951, pp. 313-323. My first contact with this model was in Samuel Goldberg, Introduction to Difference Equations (New York John Wiley and Son, 1958), pp. 103-107.

[10] If ω is the event: "optimal response pattern by employees involved in the process", one form of reponse probability function which is consistent with the behavior we postulate is

$$P(\omega) = (1 - e^{-N}) / e^{n}$$

where N is the number of previous occurrences of the present stimulus, and n is the number of periods since the present stimulus was last received.

activities: the size of the change must be determined,[11] workers must be assigned, operating procedures must be suggested, delivery schedules of input materials or subassemblies must be arranged with suppliers or subcontractors.[12] If the labor force is increased or decreased to expedite the change, out-of-pocket costs of hiring (interviewing, orientation and training, and processing additional records) and firing (revision of records, final interviews, and perhaps terminal pay or even guaranteed annual wage) must be absorbed.[13]

Having examined the three factors — technical, psychological, and administrative — which underlie the costs of changing production rates, the next topic of interest is the relative magnitudes of these costs when the changes are made in the intermediate and short run.

---

[11] The distinction between the administrative costs of determining the optimal change size, and the technical costs of time consumed in making the adjustment is perhaps tenuous. The two factors are complimentary; by intensified administrative effort, the adjustment time interval can be shortened.

[12] A constant output rate permits a constant flow of input materials and results in savings on inventory cost which otherwise would be incurred for holding materials. In addition, this steady inflow rate is less costly and more convenient to the supplier of the inputs, whether they be packing materials or subassemblies; this convenience might be rewarded by a discount, provided that the steady ordering rate is maintained, because the supplier is freed from the necessity to hold large inventories of his finished product, and can smooth his own production and requisitioning schedules. It is quite possible that the holding of inventories of the same item by different firms at different stages is a widespread waste, which might be corrected by proper attention to steady production schedules at the final stage, accompanied by steady rate discounts and agreements regarding advanced notification of increase or decrease in the ordering schedule of the final stage firm, to permit smooth adjustment at earlier stages.

[13] Further, if the assumption of stationary demands is a good approximation, changes away from the rate p will be made in full expectation that a change back to p will later be necessary.

2.1 Production Change Costs: Intermediate and Short Run

Input engineering costs are encountered in the intermediate run and short run alike. The optimal input combination will no doubt be more difficult to identify in the intermediate than in the short run, since the number of possible input combinations to be considered will be greater.

Similarly, when the intermediate run alternative is adopted, learning costs will be higher, since it will be more difficult for new employees to adjust to an entirely foreign routine than for established employees to adjust to a changed output rate. Finally, the costs of hiring and firing will cause administrative costs of production change to be higher in the intermediate run. Also, because of the smaller number of technical input combinations in the short run, the expenditure of managerial effort in calculating this optimum will be reduced.

The time path of adjustment from production rate p to $\bar{p}$ is depicted in Diagram II.[14] While the comparative static production costs are lower in the intermediate run, the dynamic costs of making the change are seen to be higher. Thus, if s_t and i_t are respectively the short-run

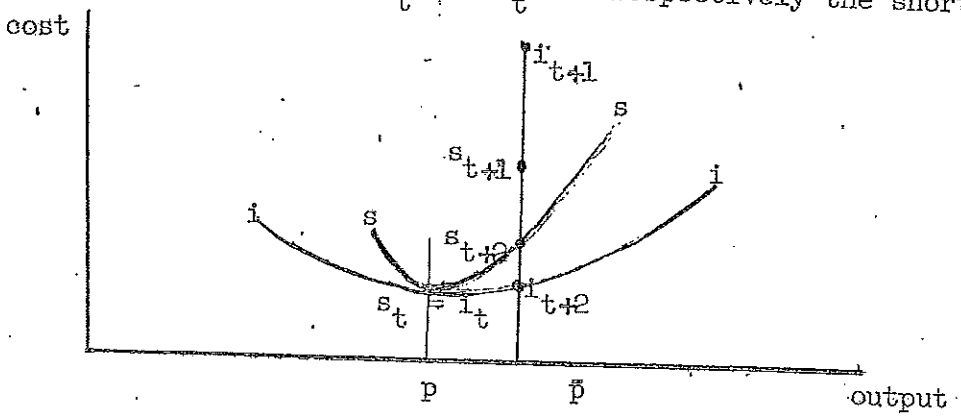

Diagram II
Adjustment from p to $\bar{p}$ in the short and intermediate run.

[14] It should not be inferred that the optimal size of $\bar{p}$ will be the same in the short run and intermediate run (see the remarks regarding evaluation of the two alternatives: infra, Section 3). They are treated as the same in the diagram to facilitate comparison.

and intermediate run cost and output, the time paths of adjustment are $s_t : s_{t+1} : s_{t+2}$ and $i_t : i_{t+1} : i_{t+2}$. As seen in (16) the distances $i_{t+1} - i_{t+2}$ and $s_{t+1} - s_{t+2}$ may depend upon the size of the change $\bar{p} - p$, depending upon whether $\bar{k}$ is positive or zero.

2.2 Costs of Change and the Number of Production Rates

The goal of the (a,b,c) policy, to control production rate changes, is rationalized on the basis of the foregoing cost description. In addition, the same arguments which were presented in favor of a constant rate of production may be applied in choosing between a small number of possible operating rates versus a large number. In Section 2, the factors which cause production change costs and determine the magnitude of these costs were identified. The purpose of this section is to show that restricting the set of possible production alternatives to three rates has the desirable effect of reducing change cost.

Consider first the effect of the limited set of alternatives upon the cost of production engineering. These were seen to stem from three factors: problems of description, of input cost changes, and of technological changes. If only one increase and one decrease from the 'normal' operating rate p are permitted, the problems of description remain resolved until technological changes or input cost changes necessitate a recomputation of the optimal size of production change. In times of technological and price stability (in which several periods may go by without perceptible change in either of these categories) the advantage of controlling the number of increases and decreases will be substantial.

The second causal factor, costs of learning, also reflect the desirability of a limited number of production alternatives. This becomes clear upon considering the learning model discussed earlier (Section 2.)

Like the costs of learning, the administrative costs can be expected to decline as practice in the process of making a change is assimilated. Given price and technological stability, the necessity to recompute the size of change can be averted entirely.

3 Effect of the Employment Alternative Upon Computation

In deciding which specific employment alternative should be adopted, the expected cost of each must be evaluated separately in the context of the (a,b,c) policy, and the less costly chosen. Thus, the loss function (50) would be evaluated twice, using the different cost parameters associated with the two employment alternatives. It is entirely conceivable that assymetries may arise: it may be more economical to move from p to $\underline{p}$ in the intermediate run, while the short run alternative may be less costly in moving from p to $\bar{p}$. Of course, unless the short run costs of increase equal the short run costs of decrease (and similarly for the costs of adjustment in the intermediate run), the parameters must be determined by minimizing some modification of (39) rather than (50), since the latter formulation of the loss function is dependent upon the assumption of cost symmetry. In this event, a function of six, rather than four, variables is encountered.

On the basis of the discussion of Sections 2 and 2.1, it is easier to appraise the assumption which is fundamental to the theorem of Section II-3.3, namely that costs of production change are equal at the upper and lower barriers, and which now must be restated: the costs of production change are equal at the barriers a and b when both changes are evaluated in the short run or in the intermediate run. When inventory falls to the level b, production is immediately increased to $\bar{p}$, but since it will later

be decreased (with probability equal to one) back to the rate p, change costs accrue from one increase and one decrease in production. Similarly, passage above a will signal a decrease of $p - \underline{p}$ units, followed by an increase of the same size later. Since one increase and one decrease is involved in both types of change, the constant elements of the production change cost function, $\underline{K}$ and $\bar{K}$, may very likely be equal, and similarly, the costs per unit of change, $\underline{k}$ and $\bar{k}$, may also be equal. Thus, any differences which arise will be the result of different magnitudes $\bar{p} - p$ and $p - \underline{p}$. There is no reason to anticipate that these production changes will be equal in size; however, this difference may be unimportant, if k is small compared to K.

This argument is not intended as prima facie justification for the assumption of change cost symmetry; indeed, the analysis can proceed whether or not the assumption is satisfied. What the argument attempts to establish is the reasonableness of the assumption in the absence of reliable empirical cost data.

CHAPTER IV

CRITERIA FOR USE OF THE (a,b,c) POLICY: CONCLUSION

1 Questions of Optimality

Three questions were raised at the time the (a,b,c) policy was defined.[1] The first of these was the optimality question, which may be phrased: what set of specific conditions is necessary and sufficient to assure that the policy is optimal? That this question is exceedingly difficult to cope with is evidenced by the experience with the s,S policy: the optimality question was raised in print shortly after the appearance of the Arrow, Harris and Marschak paper, yet nearly a decade passed before Scarf established the optimality conditions.[2]

The optimality conditions of the (a,b,c) policy will not be established in this study; rather, some questions will be answered which impinge upon the suitability (a less exalted property than optimality) of the (a,b,c) policy for decisionmaking. Chapter III argues that some sort of policy capable of controlling the frequency of production rate changes may be indicated in a wide variety of circumstances, and Chapter II demonstrates that the (a,b,c) policy is a device whereby this can be accomplished. Furthermore, Chapter III also establishes that a limited number of allowable production rates is frequently a desirable attribute of production policies, an attribute which is possessed by the new policy.

---

[1] Supra, Section II-2.

[2] Supra, Chapter I, footnote 25.

Several questions remain to be answered. (a) Why permit only three production rates, or alternatively, why not be satisfied with two, or even one production rate? (b) Why establish fixed inventory 'barriers' at which production rates are changed, rather than letting the known initial inventory level and production rate determine when changes will next be made? (For example, despite the fact that a change was signalled only a few periods ago, another one is indicated in this period, but in light of the recent change, why not let it 'ride a while' before responding?) (c) Even if three production rates are allowed, how do we know that the number of inventory barriers which signal the changes is best set at three?

In answering question (b), it must be remembered that the analysis deals with the steady state, and therefore the incorporation of transient influences into the decision process is impossible for computational reasons. A model incorporating the type of transient consideration typified by question (b) would be computationally impracticable, because of the number of conditional dependencies which would be involved. This sort of transient consideration is best handled on a "spur of the moment" or "seat of the pants" basis; a policy – dictated production change could be countermanded if transient circumstances so indicated. Questions (a) and (c) are rather more penetrating, and merit more extensive discussion. Question (a) is answered first.

1.1 The Optimal Number of Production Rates

Section III-2.2 was a rationalization of the contention that the allowable number of production rates should be limited, but no reasons were there given for specifying exactly three rates, as in the (a,b,c)

policy. It is not difficult to visualize circumstances under which four, five or more rates would be more appropriate than three. With a larger number of rates, greater flexibility of operation is obtained, but at the same time, costs of production change are increased. Cost reductions are ceteris paribus realized through the increase in flexibility; for example, suppose in addition to the (a,b,c) policy rules (19), a second production increase, from $\bar{p}$ to $\bar{p}_2$, is called for at some inventory level B, where $b > B \geq 0$. Then the probability of a runout can be controlled more closely, and if the penalty for runouts is large, it may be economical to specify the second increase. A move directly from p to $\bar{p}_2$ may be uneconomical because production costs $m(\bar{p}_2)$ are substantially higher than $m(\bar{p})$; the second increase is made only as a last resort, when a runout seems imminent. Similar arguments may hold in favor of another rate lower than $\underline{p}$, to be effected when inventory passes above some critically high level A, where $A > a$.

Let v represent the number of production rates specified in the policy, and c(v) be the "flexibility associated costs", those costs which decrease as the allowable number of rates is increased. Then c(v) is assumed to have these properties:

$c(v) \to \infty$ as $v \to 1$ from the right.

$c(v) \to Q$ as $v \to \infty$, $Q \geq 0$.

$c'(v) < 0$ and $c''(v) > 0$, $1 < v < \infty$.

Let I(v) represent those costs which increase with the allowable number of production rates (including production change costs). I(v) is characterized by

$I'(v) > 0$ and $I''(v) \geq 0$, $1 < v < \infty$.

Therefore $T(v) = c(v) + I(v)$ is convex, and has a minimum at $v > 1$. The

indefinite increase of costs as the number of allowable rates declines toward 1 follows from the fact that in taking convolutions of independent distributions, the means and variances are added. Thus, if the demand distribution has nonzero variance, the variance of the stationary inventory distribution will increase without bound as the number of periods increases, regardless of which single production rate is chosen. Since this variance increases without bound, the probability of perpetual stockout or indefinitely large inventory approaches 1. Unless change costs completely override the other costs of the model, this pathological behavior is to be avoided.

Next, policies which specify two production rates are considered. One rate must be greater, and the other less than mean demands; for otherwise the probability of either unbounded accumulation or decumulation approaches 1 as the number of periods considered increases, and the costs noticed in the case of one production rate will be encountered. Let the policy be

$$(56)\quad z_{t+1} = \begin{cases} \bar{p} & \begin{cases} X_t < b, & z_t = \underline{p} \\ X_t < a, & z_t = \bar{p} \end{cases} \\ \underline{p} & \begin{cases} X_t > a, & z_t = \bar{p} \\ X_t > b, & z_t = \underline{p} \end{cases} \end{cases}$$

By selecting $\underline{p}$ and $\bar{p}$ sufficiently close to mean demands $E(\xi)$, the cycle time of the two drift policy can be made greater than the cycle time of the (a,b,c) policy,[3] and the frequency of production changes thereby reduced.

---

[3] The cycle time of the policy (56) is given by $2r(1/\bar{\mu} + 1/\underline{\mu})$ where $2r = a - b$, $\bar{\mu} = \bar{p} - E(\xi)$, and $\underline{\mu} = E(\xi) - \underline{p}$. When the assumptions of Section II-3.3 hold, the cycle time of the (a,b,c) policy is given by $r(r + 1/2\bar{\mu} + 1/2\underline{\mu})$. Then, if r, $\bar{\mu}$ and $\underline{\mu}$ are the same for both policies, the cycle time of the new (a,b) policy will be greater than that of the (a,b,c) policy when $r < 3(\bar{\mu} + \underline{\mu})/\bar{\mu}\underline{\mu}$, not an unreasonable condition.

However, even if fewer production changes are realized, this is not sufficient to assure that the (a,b) policy (56) will perform in a less costly fashion than the (a,b,c) policy. Higher production costs are constantly incurred, since the minimum cost production rate p is never used, for instance. Using the same rationale, approximation methods, and checks upon limiting behavior as with the (a,b,c) policy in Sections II-3.4 and II-3.5, the loss function associated with the (a,b) policy is readily found; a rough gauge of the suitability of the (a,b,c) policy is obtained by comparing the minimum attainable value of (50) to the minimum of[4]

$$(57)\quad L(a,b) = T[\Theta(r+b)/\underline{\mu} + K\bar{\mu}\underline{\mu}/2r(\bar{\mu}+\underline{\mu}) + (k/2r+m)\bar{\mu}\underline{\mu} + v\underline{\mu}(e^{-2\bar{\mu}r}-1)/(e^{-2\bar{\mu}r}-e^{2\bar{\mu}b})]$$

where the successive terms are cost of holding, production change, production, and runout.

If the (a,b,c) policy emerges as less costly than any policy specifying two production rates, the next step is to compare it to the least costly policy involving four production rates. The form of an optimal policy using four rates is an open question, and will vary in different physical situations. However, once this policy is identified, if it is more costly than the (a,b,c) policy, it is known by the convexity of T(v) that no policy specifying more than four rates will be less costly than the (a,b,c) policy, and if a less costly policy exists, it will specify exactly three production rates. By this procedure it is possible to compare policies which specify different numbers of production rates, but leaves unsolved the problem of how to compare different policies

[4] Of course, it is first necessary to investigate whether another two-drift policy is capable of lower expected costs than (56). In deriving the loss function for the (a,b) policy, we have assumed that all costs are of the same form as was specified in Section II-3.2.

having the same number of production rates. Fortunately, when the allowable number of rates is small, the number of sensible policies of simple form is also likely to be small, so it should be possible, if equipped with a high speed computer, to exhaust the likely alternatives.

1. 2 Excess Variables and the Number of Change Signals

Question (c) of Section I concerns the optimal number of change signals to be included in the policy rules, and is perhaps the most salient specific question which can be raised regarding whether the (a,b,c) policy is the best policy of the set specifying three production rates.

The formulations of the cost relations of the (a,b,c) policy, as shown in Section II-3.2, are approximate, since they ignore the presence and influence of excess variables.[5] These excess variables can be ignored when the (a-b) range is large compared to the standard deviation of demands, but cost calculations are upset when this ratio is not sufficiently large. Due to these excesses, the random walks do not begin precisely at barriers: for example, the walk in the plus drift will begin, on the average, at a level $b - E(s_b)$, where s_b is the magnitude of the passage beyond the barrier b, the amount by which the variate X "spills over" past the barrier before the passage is detected.[6] This discrepancy has two effects upon the estimation of costs: first, the runout probability measure is understated, because the plus drift begins $E(s_b)$ units lower

[5] Supra. p. 32

[6] The precise magnitudes of the excess variables are difficult to obtain, since they will depend upon the stationary inventory distribution. For applications, approximations can be developed, e.g. by assuming the stationary inventory distribution is uniform.

than is anticipated in the approximation measure, and second, the expected duration of the plus drift process is $(c + E(s_b) - b)/\mu$, rather than $(c - b)/\mu$. The excess variable also may have disturbing effects at the barriers a and c, for similar reasons. In particular, the zero drift duration calculation is upset. If the conditions of Section II-3.3 hold, $c = (a + b)/2$ and $D(0)$ equals r^2, where $r = (a + b)/2$. With an excess of $E(s_c)$ beyond c in the plus drift, and $E(S_c)$ beyond c in the minus drift, the zero drift will begin at $c + E(s_c)$ or $c - E(S_c)$, leading to the mean duration

$$(58) \quad D(0) = r^2 - [E(S_c)^2 + E(s_c)^2]/2.$$

Because of this effect, it may be desirable to recompute the loss function (39), basing all duration and runout probability calculations on the assumption that the drifts begin at the appropriate barrier, plus or minus the appropriate mean excess, as was done in (58). In particular, by installing two signal levels, $c_a = c + E(S_c)$ and $c_b = c - E(s_c)$ in lieu of the single signal level c, it is possible to decrease the dispersion about c of the beginning points of successive zero drifts. Under the circumstance that such a move enhances the future pattern of expected costs, a three- rate policy with more than three signal levels will be preferred to the (a,b,c) policy.[7]

---

[7] By (58), the impact of excess variables upon the zero drift duration can be represented as follows: if $E(S_c)$ and $E(s_c)$ are approximately equal, and their ratio to the range r is $1/Q$, the approximation (34) overstates $D(0)$ by the fraction $Q^2/(Q^2-1)$. For ratios of 1/10 or less, $D(0)$ will be overstated by 1 per cent or less. Further, unless the excess variables at the barriers b and a are taken into account in computing $D(+)$ and $D(-)$, the effect upon the cycle time of the error in $D(0)$ will be even smaller than our estimation of the error indicates, for the overestimation of $D(0)$ is offset by underestimations of $D(+)$ and $D(-)$. Clearly, as the range a - b increases vis a vis the excess variables, their influence upon passage times rapidly becomes, in the words of Professor Feller, "small trash", and may be relegated to the limbo of $o(a-b)$.

2 Prospectus for Application

The purely formal aspects of this study have thus far been stressed almost to the complete exclusion of the interesting question of application. This dichotomy between formal and applied analysis is quite noticeable in the theory of inventory and production; however, since this study represents an attempt to generalize upon scheduling experience in a particular firm, it is appropriate to consider its applicability.[8]

Several problems which can be suppressed in the formal analysis are encountered when the question of application is raised. The first of these concerns the validity of the model: not whether the model is a faithful representation of the particular physical situation, but rather: Does the model oversimplify (or misrepresent) so that the likely result is a policy which, though optimal in the context of the model, is absurd in the context of the manufacturing firm? Several other questions of equal importance also obtrude; these must be answered in the sequence in which they are raised. (b) Must the policy developed in the formal analysis be modified to cope with factors not recognized in the model? If so, what is the best method of making and testing these modifications? (c) How should the policy be installed to assure that it will operate smoothly, with a

---

[8] It was pointed out in the preface that the Procter and Gamble Company has installed a modification of the (a,b,c) policy for manufacturing control, the modification being necessary because of joint production and other deviations from the model of Section II-2. The experience at P. and G. has thus far been a happy one. The new policy requires substantially less forecasting precision than the old policy, and difficulties incurred by stock runout or threat of stock runout were completely avoided during the first year of experience, despite a slight decrease in mean inventory levels. Most important, unplanned changes in the total production rate were kept down to the desired level of four, in contrast to the twelve (one per month) which were called for by the preceding policy. (This information was obtained from conversations with G. D. Montillon and P. S. Willard in February 1960.)

minimum of administrative attention? (d) What are the prospective savings to the firm from adopting the new policy?

The first question is not concerned with how faithfully the model incorporates every nuance of the physical system: rather, the question is whether the policy based upon the model is a suitable basis upon which to construct a production and inventory control system which will not be upset by the numerous vagaries from which we have abstracted. In an applied context, the question of optimality so conspicuous in formal studies is likely to be replaced by the humbler query: will the proposed policy operate smoothly, and provide significant savings? A confident affirmative answer to this question is likely to be sufficient grounds for adoption: should a more promising policy emerge later, it can replace the one presently under consideration; in the meantime lower cost can be enjoyed.

Adopting this position, Section II-2 is reviewed to see whether any of the assumptions of the model are vital to the suitability of the (a,b,c.) policy. Discussion of how to accomodate factors which conflict with the assumptions of the model will also serve as an answer to question (b) above.

2.1. The Basic Assumptions and Their Influence Upon Applicability

The first assumption, that the firm continuously processes one product in one factory, is essential to the analysis in its current state, but suitable modifications can be made when several distinct products are alternately produced on the same equipment. The essence of the method for handling this complication is to build up a "production run inventory" of each commodity, which will be sufficient to cover expected demands

during the part of the production run while other commodities are being produced. This production run inventory is added to the part which changes with random demand fluctuations, i.e. which is controlled by the (a,b,c) policy. Several additional decisions are imposed by the presence of several products; e.g. when the output of product A must be increased, should the output of some other product or products be reduced so total production remains constant, or should total output be increased? Rules of thumb are usually devised whereby alternatives such as these may be evaluated.

The presence of several factories in a market area does not necessarily lead to higher inventories throughout the system. The area may be allocated among the factories on the basis of expected demands, so factory y can serve part of the normal marketing area of factory x when demands at x are higher than expected and demands at y are lower than expected.

Assumption 2, that demands are stationary and independent with known distribution, is fundamental to all the foregoing analysis. In the absence of randomness and stationarity, i.e. when there are trend, cycle, or seasonal components in the demand series, a substantial portion of the rationale for the (a,b,c) policy is jeopardized. This is particularly true of the argument that the most frequently used production rate is the minimum cost rate. This may enhance the desirability of the (a,b) policy, described in Section 1.1, since that policy does not rely for justification upon production at the minimum-cost rate. The influence of these identifiable demand movements upon the suitability of the (a,b,c) policy will depend upon the flexibility of capital equipment, and the relative magnitudes of these movements

compared to the random fluctuations of demand. It may happen during a period of seasonal and cyclical upswings, that the (a,b,c) policy, if calculated in ignorance of the regular demand movements, will call for an extraordinary number of production increases at extraordinarily frequent intervals. Indeed, unless the change from p to $\bar{p}$ is sufficiently large, the system may remain out of stock. Under these circumstances, ameliorating measures can readily be devised: by installing a second-level increase $\bar{p}_2$,[9] and moving c toward the upper barrier a, control can be regained over stock runouts. However, the difficulties in specifying "optimal" responses through a complete cyclical-seasonal interaction are clear. One feasible alternative, which may be used when forecasting techniques are reliable, is the systematic accumulation of stock in anticipation of high demand, or decumulation in anticipation of low demand.

Another consideration is introduced by Assumption 2: the population distribution of random demand movements (5) can never actually be known, as we have assumed; and to assume that this distribution is of some convenient form (e.g. normal) may introduce errors. For instance, if probability significance is attached to the standard deviation of the demand sample as though the demand population were normal, probability of runout, passage times, etc. may be underestimated or overestimated. Statistical tests exist whereby such errors can be detected, and these should be utilized.

Assumption 3, that production rate changes can be effected instantly, will be more realistic if schedule changes are made without variations in

---

[9] See the discussion supra. Sec. 1.1.

the labor force. Much will depend upon the time period which is the unit of reference. If the period is one hour, the assumption of instantaneous change (or its equivalent, that the rescheduled volume can be obtained in the first period after the change signal) is highly tenuous. But if a longer period (e.g. a week) is the unit of reference, the assumption is quite plausible, unless really extensive labor force changes are required. If the delay in increasing output rates is due to delay in increasing the inflow rate of materials, a remedy is to hold sufficient materials inventories to satisfy the increased requirement while the rate of materials acquisition is being increased.

Assumption 4, that the rate of production is the only decision variable, is subject to a variety of modifications; no attempt will be made to enumerate all of these. The objectives of the (a,b,c) policy may be coordinated with other objectives; an example is the scheduling of "special offer" promotions. In anticipation of a future promotion, inventory of both regular and special product is accumulated in advance of the promotion, to avoid production overload at the time of heavy demand. In general, the impact upon production loads should be considered before varying price or promotional policy; if this is done, the policy can handle such changes in a smooth and trouble-free way.

The (a,b,c) policy is designed to accomodate a particular aspect of assumption 5, the fixed costs incurred when the production rate is changed. Unless change costs conform to this description, it is doubtful whether use of the policy would make sense -- not because it would necessarily perform badly, but because other policies [e.g. those described in equation (0) or equation (14)] would probably perform better.

Assumption 6 will hold if the policy performs successfully, for unless the stationary inventory distribution exists, inventories are either increasing or decreasing without bound, and therefore infinitely high costs are incurred.

The conclusion to be drawn is that the (a,b,c) policy merits serious consideration for application wherever fixed costs of production changes are encountered, regardless of whether the other conditions assumed in the model exist. Modifications are possible which enable the policy to cope with the various shocks and disturbances assumed out of existence in the model.

2.2. Procedure for Installation

Question (c) of Section 1 is concerned with an order of operation for installing the policy. To simplify and shorten the discussion, this question is answered as though the assumptions of the original model (Section II-2) hold. An order will be indicated in which operations necessary to assure a smooth installation of the policy might take place.

The first problem is to determine explicit functions for the various pertinent costs. For the costs of holding, production and production change, this is likely to involve the fitting of functions to data gathered by industrial engineers, while runout cost is perhaps best evaluated by asking management to determine the point at which the runout protection of a marginal unit of inventory is worth less than the cost of holding that inventory.[10]

---

[10] Whitin, op.cit., pp.220-221, provides a good discussion of factors to be considered in determining holding cost. Production cost and production change cost may be determined by industrial engineering techniques (cf. the discussion of Chapter III above). The "opportunity costing" evaluation of inventory runouts is described at length in Andrew Vazsonyi, Scientific Programming in Business and Industry (New York: John Wiley and Sons, 1958) pp. 307-315.

The second problem is to identify an analytic distribution form to describe the behavior of demand. This is a statistical problem, and the assistance of a statistician will be valuable in solving it. The remarks made in the preceding section regarding the properties of a suitable approximation are repeated: a distribution form is not a suitable approximation if it leads to significant misstatement of the passage times or runout probabilities of the loss function. Furthermore, it is desirable to obtain a distribution form for which the moment generating function can be written simply and explicitly, otherwise the analytic procedure developed in Chapter II becomes exceedingly complicated, or fails entirely.

Having obtained the requisite data, the third step is to minimize equation (50) provided the assumptions of Section II-3.3 are tenable. For this purpose it will be necessary to obtain access to a computer capable of performing this operation, such as the IBM 704, or perhaps even the IBM 650. Presumably, the code mentioned in Chapter II, footnote 17, or some other code based upon gradient methods, will be utilized to this end.

Fourth, a sensitivity analysis should be performed, to test the effect upon the expected total cost of deviations in the individual cost components from the representations devised at the first step, or the deviations in the demand distribution from the representation devised at the second step. This can easily be accomplished with the assistance of a computer: by the so-called Monte Carlo method,[11] a vast variety of simulated experience may be accumulated, which will

---

[11] For a description in the context of the present study, see P. A. P. Moran, The Theory of Storage (London, Methuen and Company, 1960) pp. 83-96.

aid in determining the effects of these possible misrepresentations. Also of interest is the response of cost to changes in individual decision variables, away from the values indicated by the analysis to yield minimum cost. In other words, a picture is obtained of the loss function in the neighborhood of the minimum cost set of decision variables. If the function in this region is reasonably flat, the policy may be installed with confidence that it will operate smoothly.

Simulation provides an inexpensive and straightforward method for testing the efficacy of any policy under varying conditions; in particular, when alternative methods are proposed for handling a contingency not recognized in the original model, each possibility may be simulated, and the best one chosen.

The last step in installation of the new system is to provide for monitoring and control. Once the four decision variables of equation (50) have been identified, the process of decision-making from period to period is completely routine: for example, the level of finished goods inventory is checked after the last shipment on Friday, and the production rate for the coming week is determined in accordance with the policy rules (19). Furthermore, once an optimal method has been established for making production increases and decreases, that method will be utilized until technology changes, input factor prices change, or some other disturbing influence assumed out of existence in our model upsets the routine. When this occurs, the decision variables must again be determined, and the evaluation of short vs. intermediate run change alternatives must again be compared. A systematic way of reacting to such changes is to recompute decision variable values at regular inter-

vals, e.g. three to six months, with extra recomputations in the event of a particularly pronounced disturbance.

2.3 Potential Gain From the (a,b,c) Policy

The potential savings involved in adoption of the (a,b,c) policy are extremely difficult to appraise, since much obviously depends upon the individual circumstances of the firm, in particular, the method of scheduling in use prior to adoption of the (a,b,c) policy.

For example, suppose that a firm had been using a policy of the type shown in equation (0), where $\hat{\xi} = E(\xi)$; α, the smoothing constant, is unspecified; and X^*, the ideal inventory, is set equal to $q\sigma_\xi$, or q standard deviations of the demand distribution $\Phi(\xi)$. In this policy, the expected inventory will be X^* units, and production changes are signalled once per period. Now suppose the businessman becomes aware of production change costs of the type described in equation (16), and decides to adopt an (a,b,c) policy. A range of $2r\sigma_\xi$ is specified between the barriers a and b, and c is set halfway between a and b. The level b is set equal to X^* in the previous policy. The production rate p is set equal to mean demand, while the rates $\bar{p}$ and $\underline{p}$ are selected on the basis of experience with the previous policy: constraints encountered in dealing with that policy will be operative in determining $\bar{p}$ and $\underline{p}$. Described in the roughest terms, the firm has elected to obtain a reduction of $[100(r^2-1)/r^2]$ percent in the frequency of production changes, at the price of an increase in average finished goods inventory equal to r standard deviations of demand. (This is an extremely modest way of appraising what is gained by the increase in average inventory. The runout probability should be decreased by the

change, some beneficial effects of reducing the allowable number of production rates should be felt, and only the time spent in the zero drift is taken into account in calculating the reduced frequency of change: the duration in the plus and minus drifts have been ignored in calculating the "cycle time", or frequency of change, which means that this frequency is overestimated.)

For a numerical example, suppose the cost of making a production change is \$50, the standard deviation of demand is 100, the value per unit of product is \$10, and the holding cost is 25 per cent of value per annum. Then the approximate gain per annum of switching to an (a,b,c) policy is given by

$$[(r^2-1)/r^2]\ 2500 - 250r,$$

and the maximum gain is realized by setting $r = \sqrt[3]{20} \sim 2.71$, at which point the gain is approximately \$1485 per year. The previous total outlay had been $2500 + 250q$ per year; thus the saving is a substantial percentage of the previous outlay: typical values of q might be between 3, involving a saving of 45%, and 6, involving a saving of 37%. On one hand these figures seem conservative, since they ignore the benefits of reduced runout probability and lower change costs through limitation of the number of possible output rates; on the other hand they are predicated on the firm operating in complete ignorance of a charge per period which is the order of ten per cent of its total annual charge for holding inventory.

The example, though unrealistically formulated and imprecisely calculated, does pinpoint a critical factor: the (a,b,c) policy, by comparison with other possible policies, operates on the philosophy that it is wise to absorb higher holding costs, in order to achieve

marked reduction in production change costs, with substantial 'fringe benefits' possibly accruing in the form of reduced runout costs and lower costs of making each individual production change. All of these factors are taken systematically into account in the formulation of the loss function of the policy, (39).

3 Conclusions

This monograph, in a narrow view, may be regarded as a special study in the new mathematical theory of inventory and production, specifically concerned with describing costs of changing production rates, (and in particular, the dependence of these costs upon two factors: the frequency of production changes, and the number of allowable production rates); and prescribing a production scheduling policy capable of coping with these costs. As a first attempt to incorporate this type of cost, it has left several problems unresolved. The optimality question has not been answered, nor have responses been prescribed for such frequently encountered problems as scheduling change lags [where $\lambda > 0$ in equation (4)], trends and seasonal variations in demands and factor prices, and accomodation of strategy for marketing. However, certain questions pertinent to the suitability (a sort of weak, tentative substitute for optimality) of the new policy are answered, in the context of the model.

The question most successfully handled in this study regards the ability to identify optimal values for the decision variables of the new policy. The T-period stationary approximation yields a function which may be optimized by numerical methods on a high-speed computer. The question of the accuracy of this approximation formulation was raised in the later discussion, and it was concluded that when the range of inventory levels which call for production at the minimum cost production rate is large compared to the standard deviation of demands, the total costs which accrue

from use of the policy are quite accurately estimated.

Like all inventory policies investigated in the literature, the (a,b,c) policy as it currently stands is not a finished tool, ready to be utilized for manufacturing control. However, by calling attention to the fixed costs of changing production, and by demonstrating that the policy is capable of coping with them, a service has been performed for the industrial programmer. Optimization of the loss function (50) should give a satisfactory approximation to the optimal operating levels in a real situation, provided the empirical cost data is accurate, and the assumptions of this model do not do violence to the existing physical conditions. Even if the assumptions grossly oversimplify, and the policy is therefore unsuitable for use in an unmodified form, the cost interactions described in this study may be taken as data when making decisions in a real situation.

Cost curves derived from the production function were borrowed from the traditional static theory of the firm, and used in a slightly modified form to develop a policy for production and inventory control. Perhaps a more significant sort of reciprocity will one day be established: ideas and techniques from inventory and production theory may be incorporated in the theory of the firm. Whitin contends that the considerations central to the newer type of analysis are, or should be, incorporated into the cost curves of the traditional analysis.[12] Inventory and production costs are, as was shown here, intimately related to the rate of output, the scheduling period, the level of inventory and other dynamic considerations, and to hope that

---

[12] Op. cit. pp. 86-87. In note 28, p. 87, Whitin says "... the level of inventory in economic purchase quantities, in safety allowances, and in economic manufacturing lot sizes were all determined through the use of marginal analysis. The costs and benefits of stocking marginal units were compared in each case." It should be pointed out that the use of marginal analysis in determining inventory levels is far different from the incorporation of inventory and production costs into the cost curves of the firm: Whitin's statement implies that the operations are equivalent.

the static marginal analysis can take these factors into account would seem highly optimistic. Through synthesis with investigations of stochastic and dynamic aspects of capital budgeting, the theory of inventory and production may some day be incorporated into a theory of the firm in which dynamic adjustments, as well as static equilibria, will be observable.

In light of the analytic difficulties presented by studies similar in scope to the present one, a general stochastic-dynamic theory of the firm may be viewed in prospect as an incredibly unwieldy and intractable instrument. The advent of high speed computation makes such a theory more feasible, not only because of the ability to perform numerical analyses of complex relations such as the loss function (50), but also because of the ability to cope with irregular events afforded by the technique of simulation.[13]

There are signs that a gradual process of cross-fertilization is under way, in which business behavior, through the efforts of optimization-conscious research personnel, is moving in the directions indicated by theories which assume profit maximization as the goal; and simultaneously, the theories are growing more realistic, and hence better able to describe and analyze the total operations of the firm, as new ideas germinated in the process of programming are assimilated by them.[14]

---

[13] For a discussion of the impact of computers upon research in the social sciences, cf. Oskar Morgenstern, "Experiment and Large Scale Computation in Economics," in O. Morgenstern (editor), Economic Activity Analysis (New York; John Wiley and Sons, 1954).

[14] This optimistic note is voiced in William J. Baumol, "Marginalist Behavior and the Demand for Cash in the Light of Operations Research Experience," Review of Economics and Statistics, Vol. XL, 1958, pp. 209-215.

LIST OF REFERENCES

Books

1. Arrow, Kenneth J., Karlin, Samuel and Scarf, Herbert, Studies In The Mathematical Theory of Inventory and Production, Stanford, California: Stanford University Press, 1958.
2. Baumol, William J., Economic Theory and Operations Analysis Englewood Cliffs, N.J.: Prentice-Hall, forthcoming.
3. Bellman, Richard, Dynamic Programming, Princeton: Princeton University Press, 1957.
4. Bush, Robert R. and Estes, William K., Studies in Mathematical Learning Theory, Stanford, California: Stanford University Press, 1959.
5. Bush, Robert R. and Mosteller, Frederick, Stochastic Models for Learning, New York: John Wiley and Sons, 1955.
6. Carlson, Sune, A Study in The Pure Theory of Production, London: P.S. King and Son, 1939.
7. Cramér, Harald, Methods of Mathematical Statistics, Princeton: Princeton University Press, 1946.
8. Goode, H. H., and Machol, Robert, System Engineering, New York: McGraw Hill, 1957.
9. Ireson, W.G. and Grant, E. L. (editors), Handbook of Industrial Engineering and Management, Englewood Cliffs, N.J.: Prentice-Hall, 1955.
10. Massé, Pierre, Les Réserves et la Régulation de l'Avenir dans la Vie Economique, Paris, Hermann. et. Cie, 1946.
11. Moran, P. A. P., The Theory of Storage, London: Methuen and Co., 1960.
12. Morgenstern, Oskar (editor), Economic Activity Analysis, New York: John Wiley and Sons, 1954.
13. Whitin, Thomson M., The Theory of Inventory Management, Princeton: Princeton University Press, Second Edition, 1957.

Articles

1. Arrow, Kenneth J., Harris, Theodore, E. and Marschak, Jacob, "Optimal Inventory Policy," Econometrica, Vol. XIX, 1951, pp. 250-272.

2. Baumol, William J., "Marginalist Behavior and the Demand for Cash in the Light of Operations Research Experience," Review of Economics and Statistics, Vol. XL, 1958, pp. 209-215.

3. Bush, Robert R. and Mosteller, Frederick, "A Mathematical Model for Simple Learning," Psychological Review, Vol. 58, 1951, pp. 313-323.

4. Chenery, Hollis, "Engineering Production Functions," Quarterly Journal of Economics, Vol. LXV, 1951, pp. 501-531.

5. Cooper, William W., "Extending the Theory of the Firm," Quarterly Journal of Economics, Vol. LXV, 1951, pp. 87-109.

6. Feller, William, "On The Integral Equation of Renewal Theory," Annals of Mathematical Statistics, Vol. XIII, 1941, pp. 243-267.

7. Hoffman, Alan and Jacobs, Walter, "Smooth Patterns of Production," Management Science, Vol. I, 1954, pp. 86-94.

8. Pinkham, Roger, "An Approach to Linear Inventory-Production Rules," Operations Research, Vol. VI, 1958, pp. 185-188.

9. Shaw, E.S., "Elements of A Theory of Inventory," Journal of Political Economy, Vol. XVIII, 1940, pp. 465-485.

10. Simon, Herbert, "Dynamic Programming Under Uncertainty With A Quadratic Criterion Function," Econometrica, Vol. XXIV, 1956, pp. 74-81.

11. Simon, Herbert, "On The Application of Servomechanism Theory in the Study of Production Control," Econometrica, Vol. XX, 1952, pp. 247-268.

12. Wald, Abraham, "Differentiation Under The Expectation Sign in the Fundamental Identity of Sequential Analysis," Annals of Mathematical Statistics, Vol. XVIII, 1946, pp. 493-497.

13. Wald, Abraham, "On Cumulative Sums of Random Variables," Annals of Mathematical Statistics, Vol. XV, 1944, pp. 283-295.

Privately Circulated Unpublished Papers

1. Box, George E.P., The Use of Statistical Method in the Elucidation of Basic Mechanisms, Princeton University, Statistical Techniques Research Group, Technical Report no. 7, October, 1957.

2. Mills, Harlan D., Stochastic Properties of Elementary Logistic Components, Econometric Research Program, Princeton University, Research Memorandum no. 9, February, 1959.

3. Roberts, Donald M., Approximations to Optimal Policies in a Dynamic Inventory Model, Applied Mathematics and Statistics Laboratory, Stanford University, Stanford California: Technical Report no. 12, July, 1959.

4. Scarf, Herbert, The Optimality of s,S Policies in the Dynamic Inventory Problem, Applied Mathematics and Statistics Laboratories, Stanford University, Stanford, California: Technical Report no. 11, April, 1959.

ABSTRACT

Most of the recent work in the stochastic theory of inventory and production may be classified in two categories: the first and largest portion of this work is concerned with analysis of ordering policies, in which fluctuations in the rate of inflow of new stock are viewed as costless; the second is concerned with analysis of production policies, in which these fluctuations are viewed as costly, since the source of new stock is a production facility which is a part of the same system. Analyses which fall into the second category have without exception considered the size of production changes as the sole determinant of cost of fluctuation: typically cost is viewed as a linear or quadratic function of the change sizes. This "production smoothing" approach has considerable virtue, particularly from the standpoint of computation.

The present study describes and analyzes a policy which is designed for use when costs of changing production are independent of the size of the change. More precisely, the cost of change is formulated as $K + k|z_t - z_{t-1}|$, where $K > 0$, and the z_i are production rates in successive periods. The fixed charge K makes it desirable to control the frequency, as well as the size, of production rate changes.

The policy whereby production change frequency is controlled comprises three inventory levels, designated $a > c > b \geq 0$, and three production rates $\bar{p} > p > \underline{p}$, with the operating rules: a passage above a signals the production rate $\underline{p}$; a passage below b signals a production rate $\bar{p}$; and a passage above or below c signals a production rate p. This policy, called the (a,b,c) policy, embodies a second unusual feature: the number of allowable production rates is restricted to three.

When taken in conjunction with a stationary, independent random demand series, the (a,b,c) policy generates a random walk identical to one first analyzed by Abraham Wald in 1944. The probability measures determined in that analysis are used in the loss function of the model; they furnish the means whereby the six policy variables a, b, c, $\bar{p}$, p, and $\underline{p}$ are optimized.

An interesting question regarding any inventory and production policy is: what are the conditions which are sufficient to assure that the policy is optimal? This question is not answered for the (a,b,c) policy; however, it is suggested that (a) if the fixed costs of production change K are

sufficiently important, then some policy which controls the frequency of change is indicated; (b) policies may be visualized which specify any number of production rates, and still control the frequency of production changes, so a method must be devised for deciding when three rates is the optimal number, a task which is easily accomplished provided certain plausible conditions hold; (c) given that three production rates is the optimal number, a method must be devised for determining when the (a,b,c) policy is the best choice from among all policies specifying three rates; this is a problem which remains to be solved.

Three factors are identified as basic to the dynamic costs of production as they are visualized in this study. These are (a) costs of determining the optimal input combination, the so-called technological costs; (b) the administrative costs; and (c) the costs of learning incurred during the time required by the labor force to master the best pattern of motor responses associated with a new production rate. It is argued that by controlling both the frequency of change and the allowable number of output rates, these costs can be minimized, along with costs usually dealt with in inventory and production analysis.

Two additional questions are raised: (a) in view of the fact that the (a,b,c) policy is developed in conjunction with a model which embodies a series of assumptions which may not hold, is the policy a suitable device for controlling production in an industrial firm? (b) Can the newer dynamic theory of inventory and production be incorporated profitably into the traditional microeconomic theory of production? Prospects of an affirmative answer to the first question are seen to be good, provided the businessman is not overly concerned with knowing that the policy which he uses is optimal; the chief obstacle to an affirmative answer to the second question is computational difficulty.