PRODUCTION PLANNING UNDER STOCHASTIC YIELD AND SUBSTITUTABLE DEMAND: APPLICATIONS IN MANUFACTURING AND SERVICE

by

Thin-Yin Leong

B.Eng.(Hons.), National University of Singapore (1981)

SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS OF THE
DEGREE OF

DOCTOR OF PHILOSOPHY
IN MANAGEMENT

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
January 1990

Signature of Author.....
MIT Sloan School of Management
January 1990

Certified by.....
Gabriel R. Bitran
Thesis Supervisor

Accepted by.....
James B. Orlin
Chair, Doctoral Program Committee

PRODUCTION PLANNING UNDER STOCHASTIC YIELD AND SUBSTITUTABLE DEMAND: APPLICATIONS IN MANUFACTURING AND SERVICE

by

Thin-Yin Leong

Submitted to the MIT Sloan School of Management
in January 1990
in partial fulfillment of the requirements
for the Degree of Doctor of Philosophy in Management.

ABSTRACT

This thesis examines three production planning problems with stochastic yield and substitutable demand. The first two problems are in the manufacturing of semi-conductor devices and the third is in the service sector. In all three problems, there are many types of customers with each demanding different products. These products' specifications may overlap. Where the specification of one product encompasses the specification of another, the second product can substitute for the first. There is also random variability in the production process.

In the first problem, we assumed that only one process is available and that substitution among products is transitive. That is, if product 1 can substitute for 2, and 2 can substitute for 3, then 1 can substitute for 3. The decisions considered in this problem are (a) how many wafers to produce, (b) how to stock the units produced, and (c) how to allocate the units in stock to products. The second problem extends the first by considering situations where a set of candidate processes is available, transitivity does not hold, and the production capacity is limited. The third problem considers hotel room reservations planning. This problem is related to the manufacturing problems: a room reserved is like a unit planned for production and a reservation cancelled is a unit produced that fails to meet specifications. However, the hotel problem is richer in many ways and therefore poses additional challenges.

In all three problems, the demand must be satisfied from available inventory at least $100\alpha\%$ of the time. This situation is faced by management in the custom semi-conductor manufacturing facility and in hotels. The problems are formulated as stochastic linear programs with service and capacity constraints. These problems have stochastic technology coefficients and are difficult to solve. We provide deterministic linear approximations and show that they are uniformly tighter under fairly general conditions. The approximations can be made exact at the expense of more computation effort. Simple heuristics that emulate the LP approximations are formulated and tested.

Thesis Supervisor:
Dr. Gabriel R. Bitran,
The Nippon Telegraph and Telephone Professor of Management Science.

ACKNOWLEDGEMENTS

I wish to express my deepest and most sincere appreciation to my advisor and mentor Professor Gabriel Bitran for his guidance, patience, and encouragement during the research of this thesis.

I would like to thank Professor Charlie Fine and Professor Richard Larson for their generous comments and constructive criticism while serving on my thesis committee. My appreciation also goes to Professor Sriram Dasu, Professor Steve Graves, Professor Prakash Mirchandandi, Anna Piccolo, Steve Gilbert, and Dirk Cattrysse for their friendship and willing assistance. I am also indebted to M/A-COM for giving me the opportunity to study their operations.

My greatest love and affection to my dear wife Alicia and fantastic son Jonathan whose companionship and love pulled me through this difficult and challenging part of my life. Finally, my gratitude to all the wonderful international friends at FOCUS Park Street Church for their fellowship and supportive prayers.

To Chew-Kit and Ah-Sai,

my father and my mother who believed the impossible for me.

"Listen to the MUSTN'TS, child,
Listen to the DON'TS
Listen to the SHOULDN'TS
The IMPOSSIBLES, the WON'TS
Listen to the NEVER HAVES
Then listen close to me--
Anything can happen, child,
ANYTHING can be."

--LISTEN TO THE MUSTN'TS by Shel Silverstein.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

This thesis examines three production planning problems with stochastic yield and substitutable demand. The first two problems are in the manufacturing of semi-conductor devices and the third is in the service sector. The problems are motivated by real applications. The diversity of application domains serves to illustrate the general applicability of the formulation and solution approaches to other problems as well. Beyond the common fundamental characteristics of stochastic yield and substitutable demand, each problem introduces additional elements that are peculiar to it.

The problems are interesting and differ from those in the existing literature in the following ways:

a. We study these problems from the perspective of planning rather than control. Typically, stochastic production management models assume that there is infinite capacity and instantaneous recourse. In situations where this is true, it is appropriate to oversee the production operation by controlling it period by period. However where the environment is capital-intensive and service-oriented, production must be planned for an extended horizon in order to respond to the dynamics and uncertainties of internal and external factors. One source of uncertainty common to all our problems is the production yield. (We realize at this point that production and yield are not clearly defined for the application in the service sector. This discussion is postponed to the section describing the problem.)

b. Our problems have multiple products (or service options) and the demand for one product may be satisfied with another under given conditions. This is referred to as substitutability of product demand. For example, microprocessor chips that are rated at 33 MHz may be offered to a customer requesting 25 MHz chips. The customer may not be informed that he is getting "better" chips but these will certainly meet his specifications. Of course, this will only be done when there is a shortage of the 25 MHz chips, as the higher rating chips should fetch a higher market price. Similar examples exist in hotels and air transportation. Some production processes can produce units with various grades of outcome. The grades may have a total ordering as in the microprocessor chip example. In general, grades or item categories are defined by the characteristics specified by customers and items that meet the specifications of a customer (or product) may be used to satisfy the corresponding demand. The complication arises when specifications overlap. In this case, it is not obvious how items should be defined and what demand substitutability means. Certainly, one form of commonality of inventory or another must be present.

c. Substitutable demand becomes an issue only when either the demand for products is stochastic or there is co-production. Co-production occurs when multiple item categories are produced simultaneously in each production run. These items are produced in some given proportions of the lot. (The proportions are not necessarily fixed or deterministic.) When there is neither stochastic demand nor co-production, the multi-item production management (with some effort) separates into multiple single-item problems. There may still be shared capacity considerations but no decisions are needed on how to

stock and allocate the inventory since the items are stocked exclusively for each product.

d. Inventory is kept as safety and seasonal stocks because of service and capacity requirements. In some situations, the units stocked may perish. An instance of this situation is the hotel reservation planning problem. Extensions are made to incorporate this feature.

We formulate the three problems as linear programs. This is unusual because stochastic problems are typically casts as dynamic or nonlinear programs. The need to include service and capacity constraints and the demand substitution feature dictate the use of stochastic linear programs. If the random variables come from a common distribution of the stable class, these stochastic linear programs can be transformed into convex nonlinear programs. In general, stochastic linear programs under service constraints are deterministic nonlinear programs. The feasible regions of these programs are usually not expressable in closed form. We explore this aspect and provide linear approximations that are uniformly tighter than the original problem. The linear approximations can be made equivalent to the original problem by enlarging the problem infinitely. We take a simple approximation and show that the results are good and adequate for implementation. Heuristics that emulate the LP approximations are presented and tested.

The thesis brings together issues related to commonality, manufacturing quality, flexibility, and customer service. It challenges the definition and usual measures of quality in situations where there are profitable demand for the "by-products" of the production process. The notion of manufacturing flexibility is further generalized to include circumstances with co-production. The aspect of flexibility hitherto

defined as the ability to switch from one product to another should now be modified to incorporate additionally the ability to control the mix of product output to match the demands of the market. We hope the work can also contribute towards understanding how product marketing (grouping, promotion, and pricing of products) can be managed in settings similar to the ones we study.

The thesis is organized in the following manner. The problems are described individually in next three chapters. Each chapter is self-contained and the reader should expect some repetition of material. Chapter 5 serves as a technical appendix for applications. It gives some background on the uniformly-tighter linear approximation approach that we devised and compares its capability and performance against existing methods. The last chapter summarizes the thesis and suggest some directions for future research.

CHAPTER 2

DETERMINISTIC APPROXIMATIONS TO CO-PRODUCTION PROBLEMS WITH SERVICE CONSTRAINTS

1. Introduction

This chapter examines multi-period multi-item production planning problems in environments with stochastic process yields and substitutable demands. The outputs of the process have characteristics that vary in a broad band covering the needs of several customers. The functional form of the products desired by different customers are the same but their performance requirements are different. These requirements may overlap such that units produced for one customer may be used selectively to fill another customer's demand. Customers' demands must be satisfied from inventory $100\alpha\%$ of the time.

Such situations are often encountered in practice. Especially notable are those in the high-volume components manufacturing and petro-chemical processing industries. The semi-conductor and electronic components sectors, in particular, are characterized by high yield variabilities, and produce products that have different specifications and applications. For example, a component part that goes into high technology applications like aerospace instruments has tighter specification requirements than a similar part that is used in consumer products.

The units produced by the manufacturing process can be classified into a finite number of item categories according to the ranges of their specified characteristics. The total yield rate of the manufacturing process is probabilistic. Hence the percentage of acceptable units and the

relative proportions of items in each production lot can be different from run to run. The variations of the proportions among the items are correlated. The units are classified into items to simplify inventory management. The demand for products from customers is met by selecting the items that conform to the needed requirements. The requirements met by one item may also be satisfied by items that are defined by more stringent specifications. In this way, product demands are substitutable.

This chapter is based on a study performed at a custom semi-conductor manufacturing facility. Current practice at this facility does not distinguish items from products. Production runs are made to order because of the large number of product configurations. The chapter is organized as follows. The literature is briefly reviewed in section 2, followed by the detail problem description and model assumptions. The model formulation and analytical results are presented in section 4. Heuristics motivated by the analyses are described in section 5. The next section reports computational results and comments on implications of the results. The chapter ends with a summary and conclusions.

2. Literature Review

The general class of problems studied in this chapter was proposed by Bitran and Dasu [1989]. They identified a class of problems with multiple items, stochastic yields, and, more importantly, interchangeability of items to satisfy customers' demand. They framed a multi-period model with dynamic deterministic demand; production, shortage and holding costs; and product substitution structure. Drawing from the insights of the two period problem, a class of heuristics was provided for solving the multi-period problem with no capacity constraint.

Until recently, stochastic yield problems have received little attention in the literature. Whenever uncertainty is incorporated in the models, it is usually related to demand variability. Even these have certain peculiarities. Production planning problems with uncertainties usually assume that production capacity is unconstrained. This point was highlighted in [Bitran and Yanasse 1984]. Problems of this type have been thoroughly investigated in the field of inventory control. The production/inventory management literature splits into the two main streams: 1)capacitated problems with deterministic demands or 2)stochastic demands and/or yields problems with no capacity constraint.

Papers that studied yield-related problems include [Shih 1980]; [Karmarkar and Lin 1986]; [Mazzola, McCoy, and Wagner 1987]; [Moinzadeh and Lee 1987]; [Lee and Yano 1988]; [Gerchak, Vickson, and Parlar 1988]; and [Henig and Gerchak 1989]. All these problems focussed on the single item case. Yano and Lee [1989] review the lot-sizing problem when the yields are random. They reported finding little research done on multi-period problems. The measure of performance in most of the papers, the authors encountered, seek to minimize expected costs and very few have constraints on measures of service. The latter, it seems, is because their inclusion make the problem intractable rather than being irrelevant in the problem context.

Multi-item models usually consider decisions related to the production of items one at a time or in coordination. The decision-makers, in these problems, decide how much of each item to produce. Deuermeyer and Pierskalla [1978] studied processes with co-production; that is, multiple products produced simultaneously or products with by-products. They made no distinction between items and products since it did not matter in their

instance. Deuermeyer and Pierskalla [1978] consider two items and two processes. One of the two processes makes two items simultaneously, with fixed item proportions while the other can produce one given item. The model can be generalized to m processes and n items. The products' demand is stochastic with no substitution allowed and there is no capacity constraint.

Almost all of the stochastic production planning/inventory control models have penalties for product shortages. Managerially, it is sometimes difficult to quantify what the shortage costs comprise as well as their magnitudes relative to other costs. In most instances, the production facilities are evaluated on their ability to meet demand. Hence, it is more appropriate, in these instances, to model directly the service requirements. Chance constraints are often used for this purpose. Bitran and Yanasse [1984] provided deterministic approximations to the production problem with stochastic demands. Service constraints were used in place of shortage costs. The service constraints were formulated as chance constraints and were converted into their deterministic equivalent. The problem was approximated by a deterministic linear program. The authors provided parametric relative bounds for their approximations. The relative errors are small for probability distributions commonly encountered in practice.

Our model generalizes Bitran and Dasu's [1989] model to T periods and multiple items with a general product demand substitution structure. In place of shortage costs, we introduced service constraints. The approach we take follows from the work of Bitran and Yanasse [1984]. In contrast, we have uncertainty in the yield rates with given demands whereas they assumed fixed yield rates with stochastic demands. Our problem assumes the

production of multiple items but this differs from their multi-item extension in that we have co-production of the items and our products' demand is substitutable. As in [Bitran and Yanasse 1984], we use Jensen's inequality to provide relative error bounds. For a more complete bibliography of previous studies and related problems, see [Bitran and Dasu 1989], [Bitran and Yanasee 1984], and [Yano and Lee 1989].

3. Problem Description and Model Assumptions

In studying the co-production problem with stochastic yield we encountered two types of management decisions: process-product structuring decisions and production planning decisions. A production process may be set for a specific product. However, because of variation in the output characteristics, by-products, for which there may be demand, will be produced. Hence, instead of having processes specified for each individual product, a sub-set of processes can be identified with each process targeted towards a group of products. Each process produces a subset of items. The items are used to satisfy products' demand. The chain relationship is shown below. The first set of decisions consists of determining what processes to select and what products are covered by each. These higher level decisions will be addressed in the chapter 3.

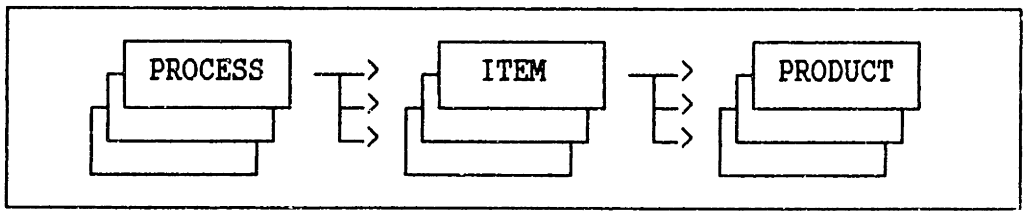


Figure 1. Process-Item-Product Chain Relationship.

For a given sub-set of products, and their pre-selected process, the production planning decisions are: 1)how much to produce and 2)how to

allocate the inventory of items to the products. We consider, in this chapter, the production planning problem under the following assumptions:

ASSUMPTIONS

a. A multi-period model with finite planning horizon. Decisions are made at the beginning of each period. Production is instantaneous or has a leadtime of a finite number of periods. The demands are deterministic and dynamic. Without loss of generality, there are no initial inventories of items. Shortages are backordered. Service requirements for meeting each product's demand are given. These are expressed as meeting or exceeding given probabilities of satisfying demand.

b. Holding and production costs are incurred in each period. All cost functions are proportional to the number of units and have the same constants of proportionality for each period. Shortages are not explicitly penalized. Undesired units may be sold for a small salvage value and revenue from this source is assumed to be negligible. Because of the above, to maximize profit we need only minimize the total cost.

c. The joint yield rate probability density function (pdf) of the items is not restricted to any type and is independent of the size of the production lot. In this way, the number of units obtained for each item is given by the product of the yield rate of the item and the production lot size. The production process is pre-selected and it has a stationary joint pdf for each period.

d. The products' demand substitution structure is known. The substitution structure allows only uni-directional (down-grading) substitution and the product substitution relations are transitive. We denote by i -> j, if item i can substitute item j. Transitive substitution means that if i -> j and j -> k, then i -> k.

4. Model Formulation and Analytical Results

Following a list of notation, we characterize the substitution structure of products. Linear programming formulations are presented next, followed by approximate deterministic equivalents.

NOTATIONS

n, T: Number of products and length of planning horizon.

A(i): Set of all products downgradeable to product i, for i=1,..,n. That is, $j \in A(i)$ implies that any item deliverable as product j, can also be delivered to the customers as product i. We say that, j is Above i in the product substitution hierarchy.

AU(i): Aggregate i, the set of all products in A(i)Ui. AU(i)=A(i)Ui.

d_{it}: Net demand of product i in period t.

D_{it}: Net demand of aggregate i in period t.

q_{it}: Yield rate of item i in period t.

We assume that, for each product i, there exists a corresponding item i that can be used directly to satisfy its demands. The yield rate of item i is the fraction of a production run that can be used for product i but not by any other product in A(i). By this definition, the yield rate of items can be very small when there are many products that have almost similar specifications. In our formulations we are interested in the sum of the yield rates of items that can be used for product i.

p_{it}: Sum of yield rates of items that can be used for product i in period t and $p_{it} = \Sigma_{k \in A(i)Ui} q_{kt}$.

f(x;y): Pdf of random variable (r.v.) x evaluated at y.

F(x;y): Cumulative density function of r.v. x evaluated at y.

Prob(.): Probability of the event argument.

$E(.)$: Expectation function.

h, c: Unit holding and unit production costs.

α: Probability target for meeting demand. (Typically, α is close to 1.)

N_t: Total number of units to be produced in period t.

I_{it}: Net quantity of items available for product i at the end of period t.

$J_{i,t}$: Net quantity of item i at the end of period t.

J_{it}^+: Inventory of item i at the end of period t. $J_{it}^+ = \text{Max}(0, J_{it})$.

J_{it}^-: Backorder of product i at the end of period t. $J_{it}^- = \text{Max}(0, -J_{it})$.

Additional notation is introduced when appropriate.

SUBSTITUTION STRUCTURE

We represent the product substitution structure by a directed graph G(V,E). The following algorithm is proposed for constructing G(V,E).

Algorithm STRUCTURE

Step 1 [Subroutine CONSTRUCT]. Construct a directed graph G(V',E'), with each product represented by a vertex in V'. We add a directed edge (i,j) if product i can substitute product j. That is i -> j <=> (i,j) ε E'.

Step 2 [Subroutine LABEL]. Re-label the graph G(V',E') with vertex labels i=1,..,n' such that for every (i, j) ε E', i < j. In this way, i -> j => i < j. Remove any cycles, discovered during the labeling process, by combining the vertices in the each cycle into a single vertex. Let the resulting number of vertices and the vertex set be denoted by n and V respectively. For each vertex i of the re-labeled graph, construct the sets A(i), for i=1,..,n.

Step 3 [Subroutine REDUCE]. Reduce the number of edges in the directed graph G(V,E') to give G(V,E) as follows:

```
SET E = E'
FOR i=1 to n; j ε A(i); k ε A(j)
  Remove (k,i) from E if (j,i) ε E
NEXT k,j,i.        ▪
```

The algorithm STRUCTURE is justified by the theorems that follow. The proofs of some lemmas and theorems are omitted to keep this manuscript within acceptable length for publication.

Theorem 1: If the product substitution relation is transitive, then the graph G(V,E) representing the substitution structure is acyclic. ▪

Theorem 2: REDUCE preserves the product substitution structure. ▪

The subroutine REDUCE eliminates the superfluous direct downgrading relations in favor of a simpler structure G(V,E) that has fewer arcs. Figure 2 below illustrates, for a simple case, the effect of REDUCE.

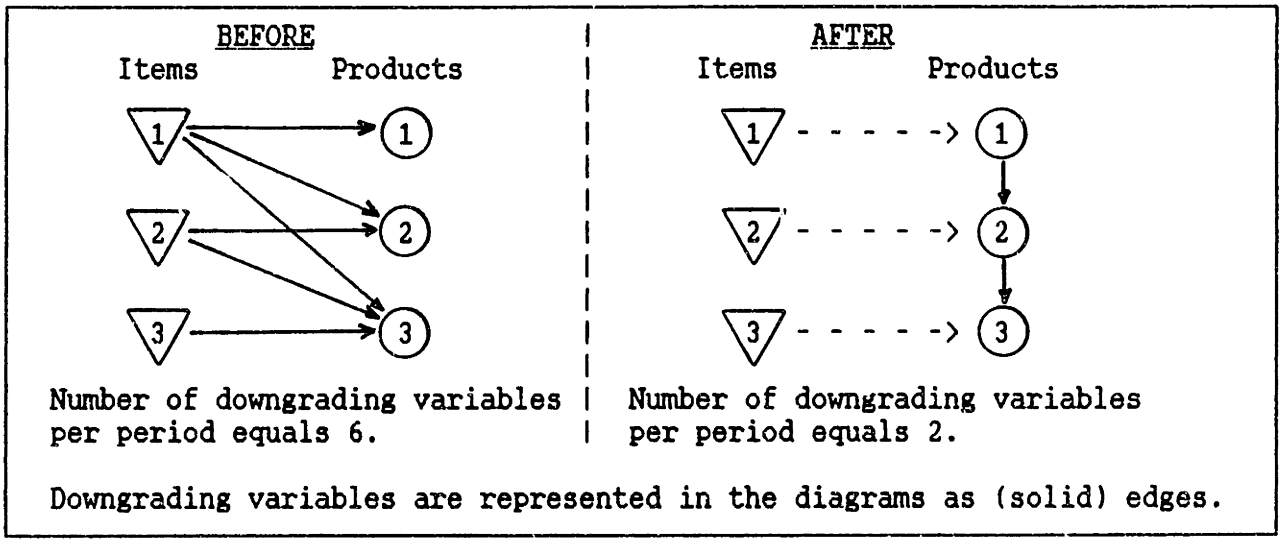
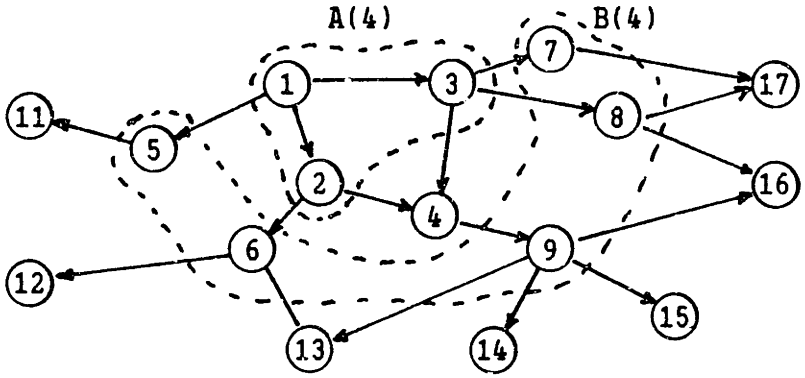


Figure 2. Effect of REDUCE on the State-Space.

Let W_{ijt} denotes the number of units to be downgraded from the inventory designated for aggregate i to product j in period t. In turn the inventory now available to product j may be used to satisfy the demand for product j or further downgraded to other products, creating a cascading effect. We define a(i) as the set of products directly downgradeable to product i. Interpreting from G(V,E), k ε a(i) <=> (k,i) ε E. Similarly,

b(i) as the set of products directly downgradeable from product i, i.e. j ε b(i) <=> (i,j) ε E. In this way, k is directly above (downgradeable to) i and j is directly below (downgradeable from) i in the product substitution hierarchy. Also, we define B(i) as the set of all products outside of AU(i) that are directly below (downgradeable from) some k ε AU(i). Figure 3 below illustrates the definitions of A(.), B(.), a(.), and b(.). (In the remainder of the chapter, these quantities refer to the graph G(V,E).)



A(4) = {1,2,3}; B(4) = {5,6,7,8,9}; a(4) = {2,3}; b(4) = {9}

Figure 3. A Product Substitution Graph.

MODEL - PROBLEM FORMULATION

(SPI)

$$Z_{SPI} = \text{Min } E(h \, \Sigma^{n}_{i=1}\Sigma^{T}_{t=1} \, J_{it}^{+} + c \, \Sigma^{T}_{t=1} \, N_t)$$

subject to

$$J_{it} = J_{i,t-1} + q_{it}N_t + \Sigma_{k\varepsilon a(i)} \, W_{kit} - \Sigma_{j\varepsilon b(i)} \, W_{ijt} - d_{it}, \qquad (1)$$

$$i=1,..,n; \; t=1,..,T$$

$$\text{Prob}(J_{it} > 0, \; i=1,..,n, \; t=1,..,T) \geq \alpha$$

$$N_t, \; W_{ijt} \geq 0, \qquad (i,j) \in E; \; t=1,..,T.$$

The first set of constraints represents the inventory balance equations, the second is the joint chance constraint on service, and the final set is the non-negativity condition on the production and downgrading quantities. The service constraint means that the probability of one or more shortages is less than or equal to 100(1-α)%. This problem is hard to

solve because of the presence of correlations among the yield rates of items. We propose to solve an alternative equivalent problem (SP), focusing on aggregates quantities rather than individual products.

(SP)

$$Z_{SP} = \text{Min } E(h\ \Sigma^{n}_{i=1}\Sigma^{T}_{t=1}\ J_{it}^{+} + c\ \Sigma^{T}_{t=1}\ N_t)$$

subject to

$$I_{it} = I_{i,t-1} + p_{it}N_t - \Sigma_{j\epsilon B(i)}\ W_{ijt} - D_{it}, \quad i=1,..,n;\ t=1,..,T \qquad (2)$$

$$\text{Prob}(I_{it} > 0) \geq \alpha, \qquad i=1,..,n;\ t=1,..,T$$

$$N_t,\ W_{ijt} \geq 0, \qquad (i,j)\ \epsilon\ E;\ t=1,..,T$$

where $I_{it} = \Sigma_{k\epsilon AU(i)}J_{kt}$, $p_{it} = \Sigma_{k\epsilon AU(i)}q_{kt}$, and $D_{it} = \Sigma_{k\epsilon AU(i)}d_{kt}$ are the aggregate variables. The number of downgrading terms in (2) is reduced because some of the 'downgrading from' and the 'downgrading to' terms cancel each other.

From (1) and (2) and noting that initial inventories are zero, we get

$$J_{it} = \Sigma^{t}_{\tau=1}\ (q_{i\tau}\ N_\tau + \Sigma_{k\epsilon a(i)}\ W_{ki\tau} - \Sigma_{j\epsilon b(i)}\ W_{ij\tau} - d_{i\tau}), \text{ and} \qquad (3)$$

$$I_{it} = \Sigma^{t}_{\tau=1}\ (p_{i\tau}\ N_\tau - \Sigma_{j\epsilon B(i)}\ W_{ij\tau} - D_{i\tau}). \qquad (4)$$

Theorem 3: (SPI) and (SP) are equivalent.

Proof:

(=>) $\text{Prob}(J_{it}>0,\ i=1,..,n,\ t=1,..,T) \geq \alpha$ => for any i and t $\text{Prob}(J_{kt}>0,\ k\epsilon A(i)Ui) \geq \alpha$. By definition $I_{it} = \Sigma_{k\epsilon AU(i)}J_{kt}$. Hence $\text{Prob}(I_{it}>0) \geq \alpha$ for any i and t.

(<=) For any i and t, we know that $\text{Prob}(I_{kt} > 0) \geq \alpha$ for $k\ \epsilon\ A(i)$. For those $k\ \epsilon\ A(i)$ such that $\text{Prob}(I_{kt} > 0) > \alpha$, we can downgrade some of their units to product i until $\text{Prob}(I_{kt} > 0) = \alpha$. Hence we can make $\text{Prob}(I_{kt} > 0) = \alpha$ for all $k\ \epsilon\ A(i)$ without changing the objective value. But $\text{Prob}(I_{it} > 0)$ can only increase with downgrading from above. Since $\text{Prob}(I_{it} > 0) \geq \alpha$, $\text{Prob}(I_{kt} > 0) = \alpha$ for all $k\ \epsilon\ A(i)$, and $I_{it} = \Sigma_{k\epsilon A(i)Ui}J_{kt}$, hence $\text{Prob}(J_{it} > 0) \geq \alpha$. Therefore, (SP) is equivalent to (SPI). ∎

We re-write (SP) as follows:

(SP)

$$Z_{SP} = \text{Min } E(h \Sigma^{n}_{i=1}\Sigma^{T}_{t=1} [\Sigma^{t}_{\tau=1} (q_{i\tau}N_{\tau} + \Sigma_{k\epsilon a(i)} W_{ki\tau} - \Sigma_{j\epsilon b(i)} W_{ij\tau} - d_{i\tau})]^{+} + c \Sigma^{T}_{t=1} N_{t})$$

subject to

$$\text{Prob}(\Sigma^{t}_{\tau=1}(p_{i\tau}N_{\tau} - \Sigma_{j\epsilon B(i)} W_{ij\tau} - D_{i\tau}) > 0) \geq \alpha, \qquad (5)$$

$$i=1,..,n;\ t=1,..,T$$

$$N_t,\ W_{ijt} \geq 0, \qquad (i,j) \in E;\ t=1,..,T.$$

The variables I_{it} and J_{it} are replaced by the right-hand-side of equations (3) and (4). We will refer to the feasible region of (SP) as G. With the joint pdf of q_{it} given, the $[\Sigma^{t}_{\tau=1} (q_{i\tau}N_{\tau} + \Sigma_{k\epsilon a(i)} W_{ki\tau} - \Sigma_{j\epsilon b(i)} W_{ij\tau} - d_{i\tau})]^{+}$ term in the objective function of (SP) can be more explicitly written as

$$\int_{\Sigma^{t}_{\tau=1} (\Sigma_{j\epsilon b(i)} W_{ij\tau} + d_{i\tau} - \Sigma_{k\epsilon a(i)} W_{ki\tau})}^{\Sigma^{t}_{\tau=1} N_{\tau}} [y - \Sigma^{t}_{\tau=1}(\Sigma_{j\epsilon b(i)} W_{ij\tau} + d_{i\tau} - \Sigma_{k\epsilon a(i)} W_{ki\tau})].f(\Sigma^{t}_{\tau=1} q_{i\tau}N_{\tau}; y)dy \ .$$

We have used for our objective function the expected value of the sum of the holding and production costs. This is not unreasonable under most situations. Other types of functions may be used to reflect risk preferences. Examples of these include the V-type and P-type formulations as proposed by Charnes and Cooper [1963] as opposed to the E-type that is used here. We will assume that the feasible region defined by constraint (5), for each i and t, is convex. That implies that G is convex. The results of Monte-Carlo simulations, under the conditions of our test problems, indicate that this is a reasonable assumption for α close to 1.

For a planning horizon of more than two periods, (SP) is difficult to solve since the yield rates q_{it} are not known beforehand. Without the prior knowledge of q_{it}, it is not possible to guarantee that any solution for the whole horizon, will be feasible after the first period. As such, most

stochastic programming problems in the literature are solved for one period at a time but may include as input, the demand of at most one period into the future. When there are seasonal demand fluctuations and limited capacity, the problem becomes even harder to solve. Hence, the need to assume that the capacity is not constrained in earlier studies.

As a step towards solving (SP), we propose a few approximations. Each of these approximations redefines the feasible region. The objective function remains the same as in (SP). We will provide the motivation and insight into each of these approximations. These alternative problems are still not solvable by standard linear programming codes because of the stochastic terms in the objective function. Deterministic approximations are then obtained for each of these formulations.

APPROXIMATIONS TO (SP)

We now focus on equations (5), the chance constraints in (SP). Since N_t, $t=1,..,T$ are our decision variables, we cannot apriori know the distribution of $\Sigma^t_{\tau=1} p_{i\tau} N_\tau$. An approximation for the constraint at period t, that is often made, is to assume that the yield rates of all periods except the latest one are equal to their expected value. This reduces the number of random variables in each constraint to one, making the problem tractable.

For each service constraint (5) for period t, we let

$$\text{for } 1 \le \tau \le t, \quad p_{i\tau} = \begin{cases} E(p_i) & ;\tau=1,..,t-1 \\ p_{it} & ;\tau=t. \end{cases}$$

The constraint (5) in period t becomes $\text{Prob}(p_{it}N_t + \Sigma^{t-1}_{\tau=1} E(p_i)N_\tau - \Sigma^t_{\tau=1}(\Sigma_{j\in B(i)} W_{ij\tau} + D_{i\tau})) > 0) \ge \alpha$ and results in:

(SP1)

$$Z_{SP1} = \text{Min } E(h\ \Sigma^{n}_{i=1}\Sigma^{T}_{t=1}\ [\Sigma^{t}_{\tau=1}\ (q_{i\tau}N_{\tau} + \Sigma_{k\epsilon a(i)}W_{ki\tau} - \Sigma_{j\epsilon b(i)}W_{ij\tau} - d_{i\tau})]^{+} + c\ \Sigma^{T}_{t=1}\ N_{t})$$

subject to

$$\phi_i(1)N_t + \Sigma^{t-1}_{\tau=1}E(p_i)N_{\tau} - \Sigma^{t}_{\tau=1}\Sigma_{j\epsilon B(i)}W_{ij\tau} \geq \Sigma^{t}_{\tau=1}D_{i\tau}, \qquad i=1,..,n;\ t=1,..,T \qquad (5.1)$$

$$N_t,\ W_{ijt} \geq 0, \qquad (i,j)\ \epsilon\ E;\ t=1,..,T$$

where $\phi_i(S) = F^{-1}(\Sigma^{S}_{s=1}p_{is};1-\alpha)$ and $\phi_i(S)$ can be interpreted as the S periods $(1-\alpha)$ fractile for items good for product i. The one period $(1-\alpha)$ fractile is the yield rate that will be exceeded with probability α. The s periods $(1-\alpha)$ fractile is the yield rate that will be exceeded with probability α if the production quantities of all the periods are equal. For simplicity of notation, we let $\Sigma^{t-1}_{\tau=1}E(p_i)N_{\tau} = 0$ for t=1. We will refer to the feasible region defined by the problem (SP1) above as G_1.

For the second approximation, in each service constraint (5), we let $p_{it}=p_i$. Here, it is as if the yield rates for each i are correlated across all the periods. With some algebraic manipulations, another approximation results. We refer to the feasible region of (SP2) below by G_2.

(SP2)

$$Z_{SP2} = \text{Min } E(h\ \Sigma^{n}_{i=1}\Sigma^{T}_{t=1}\ [\Sigma^{t}_{\tau=1}(q_{i\tau}N_{\tau} + \Sigma_{k\epsilon a(i)}W_{ki\tau} - \Sigma_{j\epsilon b(i)}W_{ij\tau} - d_{i\tau})]^{+} + c\ \Sigma^{T}_{t=1}N_{t})$$

subject to

$$\phi_i(1)\ \Sigma^{T}_{\tau=1}N_{\tau} - \Sigma^{t}_{\tau=1}\Sigma_{j\epsilon B(i)}W_{ij\tau} \geq \Sigma^{t}_{\tau=1}D_{i\tau},\ i=1,..,n; \quad t=1,..,T \qquad (5.2)$$

$$N_t,\ W_{ijt} \geq 0, \qquad (i,j)\ \epsilon\ E;\ t=1,..,T.$$

Another approach to make the random variable $\Sigma^{t}_{\tau=1}p_{i\tau}N_{\tau}$ tractable is to approximate each N_{τ}, $\tau=1,..,t$ by $\tilde{N}_t$ where $\tilde{N}_t = \Sigma^{t}_{\tau=1}N_{\tau}/t$. This implies that $\Sigma^{t}_{\tau=1}p_{i\tau}N_{\tau} \approx \tilde{N}_t\ \Sigma^{t}_{\tau=1}p_{i\tau} = (\Sigma^{t}_{s=1}N_s)(\Sigma^{t}_{\tau=1}p_{i\tau})/t$. Substituting in (5) and simplifying we obtain,

(SP3)

$$Z_{SP3} = \text{Min } E(h\ \Sigma^{n}_{i=1}\Sigma^{T}_{t=1}\ [\Sigma^{t}_{\tau=1}(q_{i\tau}N_{\tau} + \Sigma_{k\epsilon a(i)}W_{ki\tau} - \Sigma_{j\epsilon b(i)}W_{ij\tau} - d_{i\tau})]^{+} + c\ \Sigma^{T}_{t=1}\ N_t)$$

subject to

$$\phi_i(t)/t\ \Sigma^{t}_{\tau=1}N_{\tau} - \Sigma^{t}_{\tau=1}\Sigma_{j\epsilon B(i)}W_{ij\tau} \geq \Sigma^{t}_{\tau=1}D_{i\tau},\ i=1,..,n;\ t=1,..,T \quad (5.3)$$

$$N_t,\ W_{ijt} \geq 0, \qquad (i,j)\ \epsilon\ E;\ t=1,..,T.$$

We call the feasible region of this problem, G_3.

In our final approximation, we replace each chance constraint (5) by a set of K(t) linear inequalities. The linear inequalities are formed such that their extreme points are points at which selected rays from the origin intersect the lower boundary of (5). The selected rays used in (SP4) are the axes of N_t, t=1,..,T and rays in the center of the cones formed by subsets of these rays.

(SP4)

$$Z_{SP4} = \text{Min } E(h\ \Sigma^{n}_{i=1}\Sigma^{T}_{t=1}\ [\Sigma^{t}_{\tau=1}\ (q_{i\tau}N_{\tau} + \Sigma_{k\epsilon a(i)}W_{ki\tau} - \Sigma_{j\epsilon b(i)}W_{ij\tau} - d_{i\tau})]^{+} + c\ \Sigma^{T}_{t=1}\ N_t)$$

subject to

$$\Omega_{i1k}.N_1 + \ldots + \Omega_{itk}.N_t - \Sigma^{t}_{\tau=1}\Sigma_{j\epsilon B(i)}W_{ij\tau} \geq \Sigma^{t}_{\tau=1}D_{i\tau}, \quad (5.4)$$

$$i=1,..,n;\quad t=1,..,T;\ k=1,..,K(t)$$

$$N_t,\ W_{ijt} \geq 0, \qquad (i,j)\ \epsilon\ E;\ t=1,..,T.$$

The coefficients $\Omega_{it\tau k}$, $\tau=1,..,t$ in (5.4) are obtained as follows: for any i, and

1) t=1,..3, we generate t! linear constraints by permutating t coefficients $(\phi_i(\tau)-\phi_i(\tau-1))$, $\tau=1,..,t$ against the decision variables N_τ, $\tau=1,..,t$.

   (For example for t=2 and any i, the linear constraints are $\phi_i(1)N_1 + (\phi_i(2) - \phi_i(1))N_2 - \Sigma^{2}_{\tau=1}\Sigma_{j\epsilon B(i)}W_{ij\tau} \geq \Sigma^{2}_{\tau=1}D_{i\tau}$ and $(\phi_i(2) - \phi_i(1))N_1 + \phi_i(1)N_2 - \Sigma^{2}_{\tau=1}\Sigma_{j\epsilon B(i)}W_{ij\tau} \geq \Sigma^{2}_{\tau=1}D_{i\tau}$.)

2) t=4,..,T, we generate t constraints by permutating $\phi_i(1),..,\phi_i(1)$, $(\phi_i(t)-(t-1).\phi_i(1))$ against the decision variables N_τ, $\tau=1,..,t$.

The number of linear constraints needed to approximate the service constraints (5) in G_4 is $n[T(T+1)/2 + 3]$ or $O(nT^2)$. The corresponding figure for G_1, G_2, and G_3 is nT or $O(nT)$. Observe that the feasible regions of all the formulations above do not contain the stochastic yield rate term p_{it} and are deterministic. They are, however, not necessarily equivalent to the feasible region of (SP) that they approximate.

DETERMINISTIC APPROXIMATIONS

In the approximations (SP1), (SP2), (SP3), and (SP4), the objective functions are still difficult to evaluate because of the stochastic terms q_{it} and the need to compute the positive part of the inventory term. To resolve this difficulty, we propose the following deterministic approximations to each of these problems and label them accordingly. The approach is similar to the one made in [Bitran and Yanasse 1984].

First, we consider problems

(DP+1)

$$Z_{DP+1} = \text{Min } h\ \Sigma^{n}_{i=1}\Sigma^{T}_{t=1}(\Sigma^{t}_{\tau=1}(E(q_i)N_\tau + \Sigma_{k\varepsilon a(i)}W_{ki\tau} - \Sigma_{j\varepsilon b(i)}W_{ij\tau} - d_{i\tau}))^{+} + c\ \Sigma^{T}_{t=1}\ N_t$$

subject to constraints for G_1 and

(DP1)

$$Z_{DP1} = \text{Min } h\ \Sigma^{n}_{i=1}\Sigma^{T}_{t=1}(\Sigma^{t}_{\tau=1}(E(q_i)N_\tau + \Sigma_{k\varepsilon a(i)}W_{ki\tau} - \Sigma_{j\varepsilon b(i)}W_{ij\tau} - d_{i\tau})) + c\ \Sigma^{T}_{t=1}\ N_t$$

subject to constraints for G_1.

Note that the optimal solution to (DP+1) is feasible to (DP1) and it also takes on a smaller objective function value in (DP1). Hence $Z_{DP1} \leq Z_{DP+1}$. The same conclusion is true for the other approximations which are

(DPk)

$$Z_{DPk} = \text{Min } h\ \Sigma^{n}_{i=1}\Sigma^{T}_{t=1}(\Sigma^{t}_{\tau=1}(E(q_i)N_\tau + \Sigma_{k\varepsilon a(i)}W_{ki\tau} - \Sigma_{j\varepsilon b(i)}W_{ij\tau} - d_{i\tau})) + c\ \Sigma^{T}_{t=1}\ N_t$$

subject to constraints for G_k, for $k=2,..,4$, and (DP+k) for $k=2,..,4$ similar to (DP+1).

ANALYTICAL RESULTS

Fundamental Lemma (Hillier [1967]): Assume that $g_3(N,W) \geq g(N,W) \geq g_2(N,W)$

where g_k: R^{T+a} -> R^b with $N \in R^T$, $W \in R^a$ and b is the number of constraints. Consider a solution (N,W) feasible if and only if $g(N,W) \geq 0$.

i) If $g_2(N,W) \geq 0$, then (N,W) is feasible.

ii) If (N,W) is feasible, then $g_3(N,W) \geq 0$. ▪

Thus, if $g(N,W) \geq 0$ represents the exact deterministic equivalent of the constraints, then $g_2(N,W) \geq 0$ and $g_3(N,W) \geq 0$ represent constraints that are uniformly tighter and uniformly looser than $g(N,W) \geq 0$, respectively. From here on, we use the following definition.

Define: $g_k(N,W)$ by $G_k \equiv \{(N,W) \mid g_k(N,W) \geq 0\}$, for k=1,..,4.

Lemma 1 [Sufficient Conditions for feasibility to (SP)]:

$g(N,W) \geq g_2(N,W) \geq 0$. ▪

Lemma 2 [Necessary Conditions for feasibility to (SP)]: For each $i \in \{1,..,n\}$, if $P_{i\tau}$, τ=1,..,t are independent identically distributed then $g_3(N,W) \geq g(N,W) \geq 0$.

Proof: By the definition of $\phi_i(t)$, for each i and t, $\text{Prob}(\Sigma^t_{\tau=1} P_{i\tau} > \phi_i(t)) = \alpha$. Therefore for any $N_s \geq 0$, s=1,..,t, $\text{Prob}(\Sigma^t_{\tau=1} P_{i\tau}(\Sigma^t_{s=1} N_s)/t \geq \phi_i(t)(\Sigma^t_{s=1} N_s)/t) \geq \alpha$. By symmetry and an approach similar to the proof for Jensen's inequality, we can show that for all $N_s \geq 0$, s=1,..,t, G convex, and $P_{i\tau}$ i.i.d. for τ=,1..,t, $\text{Prob}(\Sigma^t_{\tau=1} P_{i\tau} N_\tau \geq \phi_i(t)/t\ \Sigma^t_{\tau=1} N_\tau) \leq \text{Prob}(\Sigma^t_{\tau=1} P_{i\tau}(\Sigma^t_{s=1} N_s)/t \geq \phi_i(t)/t\ \Sigma^t_{\tau=1} N_\tau)$. Hence, $\text{Prob}(\Sigma^t_{\tau=1} P_{i\tau} N_\tau \geq \phi_i(t)/t\ \Sigma^t_{\tau=1} N_\tau) \leq \alpha$ with equality holding when N_τ, τ=1,..,t, are all equal. Therefore, we conclude that any $(N,W) \in G$ satisfies (5.3)▪

Lemma 3 [Uniformly Tighter Constraints(i)]: $g(N,W) \geq g_4(N,W) \geq 0$.

Proof: For each i and t, the chance constraint (5) is replaced by a set of linear constraints. The extreme points formed by the intersections of these linear contraints are feasible to the chance constraint the set replaces.

By convexity of (5), any solution in the polyhedron defined by each set of linear constraints will be feasible to the chance constraint. It follows that $G_4 \subseteq G$ and $g(N,W) \geq g_4(N,W) \geq 0$. ∎

Lemma 4 [Uniformly Tighter Constraints(ii)]: If $\phi_i(s)/s \geq \phi_i(s-1)/(s-1)$ for $s=2,..,T$ and any i then $g_4(N,W) \geq g_2(N,W) \geq 0$. ∎

Lemma 5: If $\phi_i(1) \leq E(p_i)$, then $g_1(N,W) \geq g_2(N,W) \geq 0$. ∎

Though G_1 is uniformly looser than G_2, G_1 is neither uniformly looser nor tighter than G.

Theorem 4: For $s=2,..,T$, and any i, $\phi_i(s)/s \geq \phi_i(s-1)/(s-1)$ then $g_3(N,W) \geq g(N,W) \geq g_4(N,W) \geq g_2(N,W)$. ∎

To graphically depict theorem 4, we sketch below the boundaries of the feasible regions corresponding to equations (5.1) through (5.4) for product i and t=2.

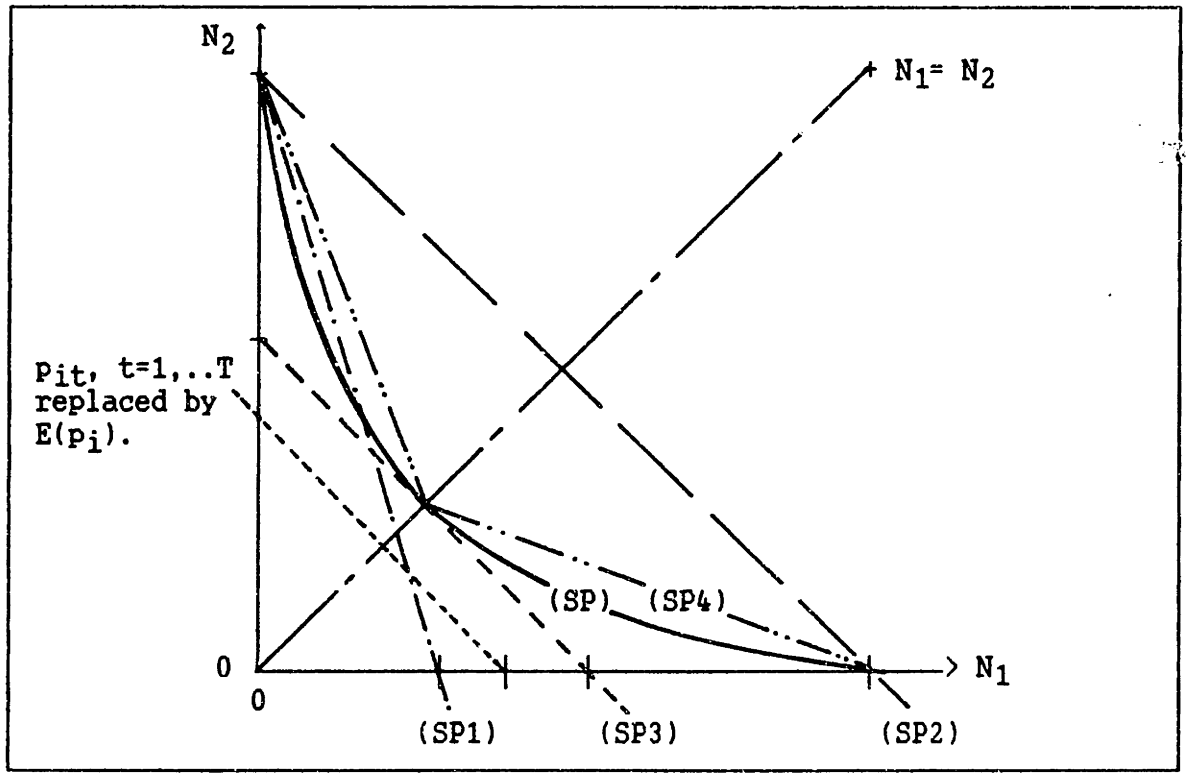


Figure 4. Boundaries of Feasible Regions.

The assumptions, for any i, $\phi_i(s)/s \geq \phi_i(s-1)/(s-1)$, for $s=2,..,T$, and $\phi_i(1) \leq E(p_i)$ are not unreasonable for most pdfs when $(1-\alpha)$ is small.

The first says that the $(1-\alpha)$ fractile of the sum of random variables after scaling for the number of terms gets larger with more terms in the sum. Plainly, it means that the risk of getting very low yield rates is less when a given production quantity is divided into more lots. This is carried forward from the conventional wisdom of not putting all the eggs in one basket. The second assumption says that the $(1-\alpha)$ fractile of a random variable is less than its expected value.

Theorem 5 [Relative Error Bounds]: Let (N^*,W^*) be the optimal solution to the deterministic approximation (DPk) under consideration. For each k, the relative error of the value of this solution to the value of the optimal solution to (SPk) is bounded above by $(ZU_k(N^*,W^*) - Z_{DPk})/Z_{DPk}$ where $ZU_k(N,W)$ is the value of any feasible solution (N,W) in (SPk).

Proof: By definition of (N^*,W^*),

$Z_{DPk} = h\ \Sigma^n_{i=1}\Sigma^T_{t=1}(\Sigma^t_{\tau=1}(E(q_i)N_t^* + \Sigma_{k\epsilon a(i)}W_{ki\tau}^* - \Sigma_{j\epsilon b(i)}W_{ij\tau}^* - d_{i\tau})) + c\ \Sigma^T_{t=1}\ N_t^*$. (N^*,W^*) optimal to (DPk) implies that it is feasible in (SPk), $Z_{SPk} \le E\{h\ \Sigma^n_{i=1}\Sigma^T_{t=1}(\Sigma^t_{\tau=1}(q_{i\tau}N_t^* + \Sigma_{k\epsilon a(i)}W_{ki\tau}^* - \Sigma_{j\epsilon b(i)}W_{ij\tau}^* - d_{i\tau}))^+ + c\ \Sigma^T_{t=1}\ N_t^*\} \equiv ZU_k(N^*,W^*)$. Note that $(\Sigma^t_{\tau=1}(q_{i\tau}N_\tau + \Sigma_{k\epsilon a(i)}W_{ki\tau} - \Sigma_{j\epsilon b(i)}W_{ij\tau} - d_{i\tau}))^+$ is convex in $(\Sigma^t_{\tau=1}q_{i\tau}N_\tau)$. Therefore by Jensen's inequality, for any (N,W),

$E\{h\Sigma^n_{i=1}\Sigma^T_{t=1}(\Sigma^t_{\tau=1}(q_{i\tau}N_\tau + \Sigma_{k\epsilon a(i)}W_{ki\tau} - \Sigma_{j\epsilon b(i)}W_{ij\tau} - d_{i\tau}))^+ + c\Sigma^T_{t=1}\ N_t\}$

$\ge h\Sigma^n_{i=1}\Sigma^T_{t=1}(\Sigma^t_{\tau=1}(E(q_i)N_\tau + \Sigma_{k\epsilon a(i)}W_{ki\tau} - \Sigma_{j\epsilon b(i)}W_{ij\tau} - d_{i\tau}))^+ + c\Sigma^T_{t=1}N_t$.

The optimal value of the left-hand side over G_k leads to $Z_{SPk} \ge Z_{DP+k}$ and hence $ZU_k(N^*,W^*) \ge Z_{SPk} \ge Z_{DP+k} \ge Z_{DPk}$. The relative error,

$REk = (Z_{SPk} - Z_{DPk})/Z_{SPk} \le (ZU_k(N^*,W^*) - Z_{DPk})/Z_{DPk}$. ∎

5. Heuristics

So far we have examined the problem with plans frozen for the whole

planning horizon. We believe these plans can be improved if they are adapted to new available information. One way of adapting is to use a rolling planning horizon. In this section, we solve linear programs (DP2), (DP3), and (DP4) to provide plans for the current period using demand information for a given horizon. We denote these as RH-SP2, RH-SP3, and RH-SP4 respectively. (DP1) was not considered because of the non-uniformity of its feasible region vis-a-vis (SP).

We next generate heuristics based on the analytical results obtained earlier. The motivation for doing this is to examine how well these simple rules derived from theoretical results can perform. If the heuristics are good, they become practical alternatives for solving the problem without relying on extensive computational power. In our heuristics, the downgrading quantities will not be computed directly. To ensure that units which have alternative uses are not double counted, we need to extend the definition of aggregates. We define the expanded aggregate i, AE(i) as equal to {i} if a(i) is empty, and {k:kϵAE(j),jϵa(i)} U {k:a(k)ϵAE(j),jϵa(i)}, otherwise. Some of the sets AE(.) may be the same. We can eliminate the redundant ones and keep only those that are distinct. The distinct AE(.) sets can be constructed using a Breadth-First Search. We redefine the sets AE(.) as AU(i), i=n+1,... . From now on we refer only to this extended set AU(i), i=1,..,2n. Depending on the product substitution structure, for n products, we can now have from n to 2n aggregates.

Two classes of heuristics were examined: heuristics with and without inventory withholding rules. We introduce three new heuristics that do not withhold inventory. In the first of these heuristics, U1, the production quantity decision mimics the deterministic approximations with one period planning horizon. (The problems (DPk) for k=1,..,4 are indistinguishable

when T=1.) For each aggregate i, we find the smallest N_i that needed to satisfy the net demand (demand less inventory plus backorders) of the aggregate. We then set the production quantity as the largest of the N_is. Product demands are met directly from the inventory of their corresponding items when possible. We examine for shortages of products in ascending order of their labels. When shortage occurs, we downgrade from their immediate predecessors in the product substitution structure, also in order of their labels, and work up the hierarchy till the shortage is resolved or no more inventory for downgrading is available. We list below the algorithm of heuristic U1 for the serial product substitution structure.

Heuristic U1

a. LET $D_{i1}^* = D_{i1} - I_{i0}$, for all i.

b. LET $N_i = D_{i1}^*/\phi_i(1)$, for all i.

$N^* = \text{Max}_i \{ 0, N_i \}$, the production quantity.

c. [The item yields q_i are realized.] Update inventory after direct assignment, $J_{i1} = q_i N^* + J_{i0} - d_{i1}$, for all i.

d. Downgrading:
FOR i=1 to n AND IF $J_{i1} < 0$
FOR j=i-1 to 1 step -1 AND IF $J_{j1} > 0$

Downgrade from j to i till i) $J_{i1} = 0$ or ii) $J_{j1} = 0$

NEXT j,i. •

The next two heuristics examine the demand of two periods and assume that the production of the next period will be the same as that of the current period. U2-SP3 mimics (DP3) and U2-SP2 mimics (DP2). The downgrading rules are as in U1. Part b of U1 is modified as follows for these two heuristics:

Heuristic U2-SP3

b. LET $N_{i1} = D_{i1}^*/\phi_i(1)$, for all i.

LET $N_{i2} = (D_{i1}^* + D_{i2})/\phi_i(2)$, for all i.

$N^* = Max_i \{ 0, N_{i1}, N_{i2} \}$, the production quantity.

Heuristic U2-SP2

b. LET $N_{i1} = D_{i1}^*/\phi_i(1)$, for all i.

LET $N_{i2} = ((D_{i1}^* + D_{i2})/\phi_i(1))/2$, for all i.

$N^* = Max_i \{ 0, N_{i1}, N_{i2} \}$, the production quantity.

The second class of heuristics holds back, under a given rule, inventory of higher order items from satisfying the demand of lower order products. The rule rations scarce higher order items so as to conserve them. This corresponds to trading-off the shortage cost of lower order items against the cost of producing more later to meet the demand of higher order items. For heuristics V, UWH01, and UWH02, the decision rule for the production quantity is the same as in U1. V is the heuristic in [Bitran and Dasu 1989]. The withholding rule in this heuristic keeps, for each downgrading source, the net product demand relative to the total demand less than or equal to its corresponding item's (1-α) fractile. Heuristics UWH01 and UWH02 are refinements of V. These two heuristics compare the relative net demands of product pairs against the ratio of their items' (1-α) fractiles. We list only the changes for each of the heuristics as follows:

Heuristic V

c. (append to end of c.)

LET $D_{i2}^{*+} = Max\ (0, D_{(i-1),2}^{*+} + d_{i,2} - J_{i1})$, for i=1,..,n where

$D_{0,2}^{*+} = 0$.

d. (replace box by)

Downgrade from j to i till
i) $J_{i1} = 0$ or ii) IF $D_{n2}^{*+} > 0$ THEN

$D_{j2}^{*+}/D_{n2}^{*+} \leq \phi_j(1)$

Update D_{k2}^{*+}, k =1,..,n
ENDIF

Heuristic UWH01

c. (append to end of c.)

LET $D_{i2}^{*+} = \text{Max}\ (0,\ D_{(i-1),2}^{*+} + d_{i,2} - J_{i1})$,

for i=1,..,n where $D_{0,2}^{*+} = 0$.

d. (replace box by)

Downgrade from j to i till
i) $J_{i1} = 0$ or ii) IF $D_{n2}^{*+} > 0$ THEN

$D_{j2}^{*+}/D_{n2}^{*+} \leq \phi_j(1)/\phi_n(1)$

Update D_{k2}^{*+}, k =1,..,n
ENDIF

Heuristic UWH02

c. (append to end of c.)

LET $D_{i2}^{*+} = \text{Max}\ (0,\ D_{(i-1),2}^{*+} + d_{i,2} - J_{i1})$, for i=1,..,n where

$D_{0,2}^{*+} = 0$.

d. (replace box by)

Downgrade from j to items k=j+1 to i
in that order of priority till
i) $J_{i1} = 0$ or ii) IF $D_{n2}^{*+} > 0$ THEN

$D_{j2}^{*+}/D_{i2}^{*+} \leq \phi_j(1)/\phi_i(1)$

Update D_{k2}^{*+}, k =1,..,n
ENDIF

6. Computational Results and Comments

The heuristics were tested on thirty test cases, each with three products having a serial substitution structure. The expected yields and

the coefficients of variation of the items relative to each other were selected so that they cover a wide variety of possible combinations. The details of the test cases are found in the appendix. We simulated the application of the heuristics for 1000 periods.

During the simulation, we calculate the average total cost per period, mean and standard deviation of production quantities per period, service levels, and statistics on inventory positions at the end of each period. Simulations for a fixed planning horizon were also done to 10 test cases randomly selected from the previous 30. Each of these was simulated for 4 periods planning horizon 1000 times. The plan was applied each time as if it was frozen for 4 periods. The upper bound on the relative errors of the deterministic approximation for the stochastic approximation are obtained using theorem 5.

RESULTS

The simulations demonstrated that the deterministic approximations under the rolling horizon perform very well. They all meet service requirements. RH-SP4 was found to perform the best. Among the LPs, RH-SP4 has the lowest average per period cost in 19 out of the 30 cases. RH-SP2 and RH-SP3 did not differ from each other at all in their performance. On the whole, RH-SP4 is 6.98% lower in cost than RH-SP3. In the best case it is 49.49% cheaper, at its worst it is 16.62% more expensive. Table 1 presents the results above. The static simulations showed that the average upper bound on the relative error of approximating (SP4) with (DP4) is about 3%.

Table 1 - LPs under Rolling horizon
(Out of 30 test cases; comparing among R-Hs.)

Methods	No. of Times Best	Average % Deviation From Best	Maximum % Deviation From Best	Average % (Max.+ %) [Max - %] Deviation From RH-SP3	Average % (Max.+ %) [Max - %] Deviation From RH-SP4
RH-SP4	19	2.17	16.62	-6.98 (16.62) [-49.49]	0.00 (0.00) [0.00]
RH-SP3	11	14.57	97.99	0.00 (0.00) [0.00]	12.36 (97.99) [-14.25]
RH-SP2	-------- SAME AS RH-SP3 ----------				

Note: Negative indicates the method is better.

From the results of the simulation, it seems advisable not to withhold inventory. The withholding of higher order items was motivated by the argument that it may be cost effective not to downgrade scarce high order items since the higher order items are relatively more difficult to produce. However, not downgrading items degrades the service performance of the lower order products. The relative scarcity of higher order items imply that the lower order items are in relative abundance. The service performance of the products corresponding to these low order items are then usually good, so withholding may not cause the service targets of these products to be violated. But if this is so, then the frequency of requests for downgrading will be so small that the additional cost incurred by downgrading, when it is needed, is negligible. Hence, it is reasonable not to restrict downgrading. This conclusion is consistent with the results in Table 2.

Table 2 - All Heuristics
(Out of 30 test cases; comparing among heuristics.)

Methods	No. of Times Best	No. of Times Second	No. of Times Violated Service Limits	When service limits are violated: Average Service Level	When service limits are violated: Worst Service Level
U1	7	9	0	-	-
U2-SP3	10	14	0	-	-
U2-SP2	12	6	0	-	-
V	7	5	12	54.93	96.00
UWH01	6	5	12	48.88	96.30
UWH02	6	5	9	48.43	36.70

Note: Best heuristics must have the lowest average per period cost as well as satisfy service limits. The number of 'best' exceeds 30 because of ties.

The main reason against using withholding heuristics is that they do not guarantee meeting service targets. Shortage probabilities for cases under withholding heuristics can be extremely high. For some of the test cases, simulation shows that under these heuristics, service requirements are violated in as many as 12 out of the 30 test cases. The average shortage probabilities among the violation cases range from 25.50% to 48.43% with the maximum service performance failing to meet demand 96.30% of the periods. The withholding heuristics do not differ very much from each other. Table 2 above presents more details.

As a whole, a myopic rule like U1 was found to do well. In fact, U1's performance was the same as RH-SP3 and RH-SP2. It appears then that, unlike RH-SP4 which was able to make use of future periods' information within its plan, RH-SP2 and RH-SP3, though both also multi-period formulations, were not able to exploit that. This does indicate that planning beyond one period is beneficial. We postulate that it will be more so when there are capacity constraints and seasonality in demand. Counting only cases that do not violate service constraints, U1 performs better than any of the other 'one period' heuristics and it will not violate service limits.

For the 'two period' heuristics, U2-SP2 is the best heuristic in 12 out of the 30 cases. This is almost twice as many times as compared to the 'one period' rules. U2-SP3, the other 'two period' rule, performed just as well with 10 firsts and 14 seconds. We now compare U1, U2-SP3 and U2-SP2 against RH-SP4, the best method. Looking at Table 3 below, it is easy to see that U1 is on the average 12.59% higher in cost than RH-SP4. U2-SP3 and U2-SP2 both perform much better with average relative deviation in cost from RH-SP4 of less than 2%. They also do better than the best method, RH-SP4, in about half of the test cases. We can conclude that the 'two-period' heuristics are much better than the 'one-period' heuristics. Also, the two 'two-period' heuristics though based on very simple rules, did almost as well as the computationally more intensive RH-SP4, a 4 period LP under rolling horizon.

Table 3 - Service Conforming Heuristics Relative to RH-SP3 and RH-SP4.
(Out of 30 test cases)

Method	No. of Times Better Than RH-SP3	WHEN WORSE Av.% Dev. From RH-SP3	WHEN WORSE Max.% Dev. From RH-SP3	ALL CASES Av.% Dev. From RH-SP3	No. of Times Better Than RH-SP4	WHEN WORSE Av.% Dev. From RH-SP4	WHEN WORSE Max.% Dev. From RH-SP4	ALL CASES Av.% Dev. From RH-SP4
U1	0	0.00	0.00	0.00	11	23.00	97.99	12.59
U2-SP3	20	5.30	25.10	-6.33	17	7.79	34.63	1.54
U2-SP2	17	6.92	33.38	-5.89	13	7.53	27.57	1.88

Another interesting result is that the coefficients of variation (COV) of production quantity of the better methods are also lower. RH-SP4's COVs are smaller than the COVs of U2-SP3 and U2-SP2. In turn U2-SP3 and U2-SP2's COV are much smaller than those of RH-SP2 and RH-SP3. In 20 out of the 30 cases, the RH-SP4's COVs are less than one half than that of RH-SP3. The average COVs are 2.69, 1.24, 2.69, 2.69, 1.89, and 1.69 for RH-SP3, RH-

SP4, RH-SP2, U1, U2-SP3, and U2-SP2 respectively. Table 4 below present the results.

Table 4 - Coefficient of Variation of Production Quantities
(Out of 30 test cases)

Methods	Av.	Std. Dev.	Max.	DEVIATIONS FROM RH-SP3 Av.	Std. Dev.	NO. OF TIMES ≥ THAN COV OF RH-SP3	RH-SP4	U2-SP3
RH-SP4	1.24	1.06	4.75	-1.45	0.72	0	0	0
RH-SP3	2.69	1.65	7.58	0.00	0.00	30	30	30
U1	2.69	1.65	7.58	0.00	0.00	30	30	30
U2-SP3	1.89	1.30	6.00	-0.81	0.50	1	30	30
U2-SP2	1.69	1.08	4.26	-1.00	0.64	0	29	15

GENERAL COMMENTS

Linear deterministic equivalents are useful and practical because sensitivity analysis can be done at no additional computational effort. This makes it easy to evaluate the cost of meeting the service requirements. Interactive-type approaches may be incorporated for adjusting the service requirements to trade-off the cost and value of the service constraints. Nonlinear deterministic equivalents and other linear deterministic equivalents have been suggested for chance-constrained problems. (See [Hillier 1967] and [Seppälä 1971].) These usually assume a particular type of pdf for the random variables. The assumption is not restrictive in most cases but does not hold for distributions that have fixed supports. Therefore, formulating the deterministic nonlinear program equivalent of our problem is already a big challenge. Also in problems where there is a large number of other linear constraints (other than those we generate to replace each chance constraint; for example, multiple resources production capacity constraints) nonlinear programming approaches become very inefficient.

Our approach is an inner linearization method. Unlike other inner linearization methods, we do not need the functions to be separable. Outer linearization approaches are usually used when nonlinear programming methods are employed. The solution to an outer linearization approximation of the problem is uniformly looser and hence may be infeasible. The gap from feasibility may be small when there are many linearization "cuts" and as mentioned in [Hillier 1967], they are "barely infeasible". The outer linearization methods are often multi-pass techniques. Our method, as presented in this chapter, solves for a planning horizon in one pass.

(DP4) is a simple version of a class of deterministic linear programs that can closely approximate the chance-constrained problem (SP). More advance, near-optimal single-pass as well as multi-pass linear programs can be constructed to approximate and solve (SP) by clever selection of rays in the construction of (SP4). We have used (SP4) in its current form for our problem and found that it is significantly better than the more common (SP2)-type approach. (For example, see [Olson and Swenseth 1988] and [Allen, Braswell, and Rao 1974].) Our approach in this chapter, increased the total number of constraints needed in our test cases from 24 (for SP2 or SP3) to 51 (for SP4).

It is interesting to note that, RH-SP4, a rolling horizon implementation of (DP4) can perform so well in a dynamic situation. Even more remarkable is that U2-SP3, a simple heuristic motivated by (DP3), differs only slightly in performance from the more sophisticated and computational more intensive RH-SP4. (U2-SP3 can also be called U2-SP4 since assuming $N_1 = N_2$ makes the second period constraints in (SP3) and (SP4) the same.)

In our computations, we have used fractiles obtained by Monte-Carlo simulations since no closed-form expression for them exists. In practice, sometimes the form as well as the values of the parameters of the joint yield distributions are not known. Historical data may be limited. In such situations, the data may be used to construct distribution-free $(1-\alpha)$ fractiles. When the form of the distribution is known, approaches similar to those in [Bache 1979] using results of Cornish and Fisher [1937] and Fisher and Cornish [1960] may be used.

In this chapter, we have assumed the capacity is unrestricted and costs constants are time-invariant. The reader will notice that these can be relaxed for the LP formulations. Heuristics can also be derived for the capacitated situation though this will require additional work. The derivation of these heuristics and evaluation of their performances, and the relaxation of other assumptions like the transitivity of substitution are examined in chapter 3.

7. Summary and Conclusions

We provided LP formulations that approximate the original problem with uniformly tighter constraints and computed, for each approximation, the corresponding optimal production plan. The uniformly tighter feature is important if planning is done infrequently since the production plan must satisfy the service constraints for the planning horizon. When planning is done every period, the approaches in this chapter provide feasible solutions even under conditions of demand seasonality and capacity constraints. Our models rely on the benefit of solving problems with more than two periods. This characteristic is particularly useful when the plans

are determined on a rolling horizon basis since they tend to change less nervously from period to period.

REFERENCES

ALLEN, F.M., R. BRASWELL, and P. RAO 1974. "Distribution-free Approximations for Chance-Constraints," Op. Res. 22, 610-621.

BACHE, N. 1979. "Approximate Percentage Points for the Distribution of a Product of Independent Positive Random Variables," Appl. Statist. 28, 158-162.

BITRAN, G.R. and H.H. YANASSE 1984. "Deterministic approximations to stochastic production problems," Op. Res. 32, 999-1018.

BITRAN, G.R. and S. DASU 1989. "Order Policies in an Environment of Stochastic Yields and Substitutable Demands," Working paper #3019-89-MS, Sloan School of Management.

CHARNES, A., and W.W. COOPER 1963. "Deterministic Equivalents for Optimizing and Satisficing under Chance Constraints," Op. Res. 11, 18-39.

CORNISH, E.A. and R.A. FISHER 1937. "Moments and cumulants in the specification of distributions," Rev. Int. Statist. Inst. 5, 307-321.

DEUERMEYER, B.L., and W.P. PIERSKALLA 1978. "A By-Product Production System with an Alternative," Mgmt. Sci. 24, 1373-1383.

FISHER, R.A. and E.A. CORNISH 1960. "The percentile points of distributions having known cumulants," Technometrics 2, 209-225.

GERCHAK, Y., R.G. VICKSON, and M. PARLAR 1988. "Periodic Review Production Models with Variable Yield and Uncertain Demand," IIE Trans. 20, 144-150.

HENIG, M. and Y. GERCHAK 1989. "The Structure of Periodic Review Policies in the Presence of Random Yield," forthcoming in Op. Res.

HILLIER, F.S. 1967. "Chance-constrained programming with 0-1 or bounded continuous decision variables," Mgmt. Sci. 14, 34-57.

KARMARKAR, U. 1987. "The multilocation multiperiod inventory problem: bounds and approximations," Mgmt. Sci. 33, 86-94.

KARMARKAR, U. and S-C. LIN 1987. "Production Planning with Uncertain Yields and Demands," Working paper, Simon Graduate School of Business Administration, U. of Rochester.

LEE, H.L. and C.A. YANO 1988. "Production control in multi-stage systems with variable yield losses," Op. Res. 36, 269-278.

LEVY, L.L. and A.H. MOORE 1967. "A Monte-Carlo technique for obtaining system reliability confidence limits from component test data," IEEE trans. Rel. R-16, 69-72.

MAZZALO, J.B., W.F. McCOY, and H.M. WAGNER 1987. "Algorithms and heuristics for variable-yield lot sizing," Nav. Res. Log 34, 67-86.

McGRILLIVRAY, A.R. and E.A. SILVER 1978. "Some concepts for inventory control under substitutable demand," INFOR 16, 47-63.

MOINZADEH, K. and H.L. LEE 1987. "A Continuous Review Inventory model with constant resupply time and defective items," Nav. Res. Log 34, 457-468.

OLSON, D.L. and S.R. SWENSETH 1988. "A Linear Approximation for Chance-Constrained Programming," J. Opl. Res. Soc. 38, 261-267.

PARLAR, M. and S.K. GOYAL 1984. "Optimal ordering decisions for two substitutable products with stochastic demands," Opsearch 21(1).

SEPPÄLÄ, Y. 1971. "Constructing Sets of Uniformly Tighter Linear approximations for a Chance Constraint," Mgmt. Sci. 17, 736-749.

SHIH, W. 1980. "Optimal Inventory Policies when stockout results from Defective Products," Int'l J. Prod. Res. 18, 677-686.

SILVER, E.A. 1976. "Establishing the Reorder Quantity when the Amount Received is Uncertain," INFOR 14, 32-39.

YANO, C.A. and H.L. LEE 1989. "Lot-Sizing with Random Yields: A Review," Technical Report 89-16, Dept. of Industrial and Operations Engineering, U. of Michigan.

APPENDIX

Test Cases

There are thirty test cases, each with three products 1, 2, and 3. Related to these products are 4 items, one for each product and the fourth for the rejects. The substitution structure is serial and transitive. That is, item 1 can be used as products 1, 2, or 3; item 2 can be used as products 2 or 3; and item 3 can only be used as product 3. The mean yield rate of each of the first three items in each problem is set L(ow), M(edium), or H(igh) relative to each other. The approximate values for L, M, and H yield rates are 0.1, 0.3, and 0.5 respectively.

We define yield rate of con-aggregate i (short for conditional aggregate) as the ratio of the sum of the yield rates of items deliverable as product i to the sum of the yield rates of items deliverable as product (i+1), for i=1, 2, 3. The coefficient of variation of each con-aggregate (CCV) is also set L, M, or H relative to each other. The con-aggregates are assumed to have Beta distributions. This is a common distribution for random variables that range between 0 and 1 and is general enough to approximate most empirical yield distributions. The $(1-\alpha)$ fractiles are generated by Monte-Carlo simulations. The test cases are set up with the parameters a and b for the distribution roughly according to the specifications outlined for each case. These cases are listed in the table below:

CCV		MEAN									
		LMH		LHL		MMM		MLM		HML	
Items		a	b	a	b	a	b	a	b	a	b
LMH	1	22	177	21	128	82	164	116	155	177	142
	2	6	7	11	2	5	3	4	3	13	2
	3	9	1	2	1	5	1	1	1	5	1
LHL	1	22	177	21	128	82	164	116	155	177	142
	2	1	1	3	1	1	1	1	1	4	1
	3	110	12	119	51	110	12	187	80	110	12
MMM	1	2	19	3	20	2	5	3	4	4	3
	2	6	7	11	2	5	3	4	3	13	2
	3	27	3	7	3	15	2	5	2	15	2
MLM	1	2	19	3	20	2	5	3	4	4	3
	2	86	108	158	26	133	66	171	128	123	15
	3	27	3	7	3	15	2	5	2	15	2
HML	1	1	6	1	3	1	1	1	1	1	1
	2	6	7	11	2	5	3	4	3	13	2
	3	110	12	119	51	110	12	187	80	110	12
HHH	1	1	4	1	3	1	1	1	1	1	1
	2	1	1	3	1	1	1	1	1	4	1
	3	5	1	1	1	5	1	1	1	5	1

Each box above contains the parameters for one test case. The total demand of all three products in each period is assumed to be uniformly distributed between 750 and 1250 units, with a mean of 1000 and a range of 500. The total demand is assigned to the 3 products according to the ratios of 3 randomly generated numbers. Unit production and holding costs are 8 and 1 respectively and α is set at 0.95.

TECHNICAL APPENDIX

PROOFS

Proof of Theorem 1: Suppose there exists a cycle in the graph G(V,E). This implies any item in the cycle can be freely substituted for any other item in the cycle. We can collapse this set of nodes into a single node. The demand for the product represented by this new node is the sum of the demands of all the products in the cycle. All arcs leading into (out of) any of the nodes in the cycle will lead into (out of) the new node. Proceeding in this way, all cycles can be reduced to nodes. ▪

Proof of Theorem 2: Consider any pair of product i and k. Suppose k -> i. Therefore by definition, $i,k \in V$, $(k,i) \in E'$, $k \in A(i)$, and $k < i$. Suppose $(k,i) \notin E$. Then by the subroutine REDUCE, there exists at least one j, such that $k < j < i$, $j \in A(i)$, and $k \in A(j)$. We focus on one of these j's denoted j_1 such that $k < j_1 < i$, $j_1 \in A(i)$, $k \in A(j_1)$, and where there does not exist any vertex l such that $l \in A(j_1)$ and $k \in A(l)$. Hence, by definition of A(.) and by the subroutine REDUCE, $(k,j_1) \in E'$ and $(k,j_1) \in E$. By repetitive application of the argument above, we can find a sequence of vertices such that (k,j_1), (j_1,j_2), (j_2,j_3),.., (j_{m-1},j_m), $(j_m, i) \in E$. Therefore, k -> i is in the product substitution structure represented by G(V,E). ▪

Proof of Lemma 1: By the definition of $\phi_i(1)$, for every i and t, $\text{Prob}(p_{it} > \phi_i(1)) = \alpha$. Therefore for any $N_t \geq 0$, t=1,..,T,

$\text{Prob}(p_{i\tau}N_\tau \geq \phi_i(1)N_\tau) \geq \alpha$; $\tau=1,..,t$.

$\Rightarrow \text{Prob}(p_{i1}N_1 \geq \phi_i(1)N_1, \ldots, p_{it}N_t \geq \phi_i(1)N_t) \geq \alpha$

$\Rightarrow \text{Prob}(\Sigma^t_{\tau=1}p_{i\tau}N_\tau \geq \phi_i(1)\Sigma^t_{\tau=1}N_\tau) \geq \alpha$

$\Rightarrow \text{Prob}(\Sigma^t_{\tau=1}p_{i\tau}N_\tau - \Sigma^t_{\tau=1}\Sigma_{j\in B(i)} W_{ij\tau} \geq \Sigma^t_{\tau=1}D_{i\tau}) \geq \alpha$

using (5.2) in the last inequality. Hence the constraints for G_2 are uniformly tighter than G. •

CHAPTER 3

CO-PRODUCTION OF SUBSTITUTABLE PRODUCTS

1. Introduction

Co-production occurs when a production run produces, simultaneously, more than one type of product. The process may be set so that it produces more of some products and less of others. Units not meeting the specifications of a target product are commonly called by-products. In our problem, it is difficult to differentiate the main product from the by-products since the products are all equally important. The outputs, in each run of production, are usually allocated among many products.

We call a set of products a family when the products in the set are by-products of each other; products form a family if they can be co-produced. In practice, the definition of a product family is usually quite clear to the operations manager. A family corresponds to the minimal set of products where inventory sharing can take place. In this chapter, we study the problem of production planning for a product family.

This problem, as well as those in [Bitran and Dasu 1989] and chapter 2 of this thesis, is based on consultancy work done with a large manufacturer of semi-conductor components. The earlier works provide the background, motivation, and review of the literature related to this class of problems. In these works, it was assumed that only one process is used and that the substitution among products is transitive. The non-transitive substitution case is also common in practice and is studied in this chapter.

Transitivity of demand substitution requires that product specifications be nested. That is, product specifications of lower order products encompass the specifications of higher order products. When the specifications of one product overlap partially with those of another, demand substitution will be non-transitive. We propose a simple approach for transforming non-transitive problems to a transitive structure.

We consider the situation where, for each family, there is more than one process. Candidate processes include current and proposed processes. Historically, new processes have been generated by making minor adjustments to an existing process to accommodate new products, or shifts in the relative demand of products. The number of candidate processes is usually much larger than the number of products, and the number of products in each family is typically less than 5. Strategically we like to identify, for each family, a small set of desirable processes. Given a set of processes, the operational problem is to determine, in each period, how much to produce and how to allocate the available inventory to the product demands.

The chapter is organized as follows. In the next section, we review briefly papers related to this work. In section 3, we describe the problem in detail and examine the characteristics of product substitution structures. In section 4, we formulate the problem as a stochastic linear program (LP). We, then, derive more tractable deterministic approximations. In section 5, we use the main properties of one approximation to suggest some heuristics. The next section reports how these heuristics perform on randomly generated test cases. The chapter ends with a summary.

2. Literature Review

A production planning problem with co-production was studied by Deuermeyer and Pierskella [1978]. Their problem considers two processes and

two products. Process A can produce products 1 and 2 in fixed proportions, and process B can produce only product 1. The product demands are stochastic, but not substitutable. Under the assumption of unlimited capacity, the authors showed that, for each period, the product demand state-space can be divided into four regions: region I, use A and B; region II, use A only; region III, use B only; and region IV, use neither. Bitran and Dasu [1989] considered the case of one process and many products. They showed that for the two products and two periods case, the optimum inventory allocation among substitutable products is determined by the relative sizes of their net demands. Net demand is defined as quantity demanded less inventory plus backorders. In chapter 2, we studied the same basic problem with service constraints instead of backorder costs.

Veinott [1965] and Topkis [1968] studied problems where customers belong to priority classes. By treating each priority class as a product, these problems are equivalent to those having several products and complete interchangeability among product demands. The authors assumed deterministic yield and stochastic demand.

Multi-item problems with shared production capacity have been studied extensively. It is usual to assume that all parameters are deterministic or to replace the stochastic parameters by deterministic estimates (e.g. their expectations). Bitran and Yanasse [1984] showed that deterministic approximations can be quite good for commonly-used distributions. It can be shown that replacing random variables with their expectations, as a general rule, is not necessarily a good approach.

A typical multi-item production planning problem separates into single-item problems if capacity is not limited. Capacitated multi-item single-period problems with stochastic demand have solution similar to that of the well-known newsvendor problem. A Lagrangian multiplier is included

in the ratio of costs to account for the shared capacity. (See, for example, [Silver and Peterson 1985] for details). The problem studied in this chapter, even without capacity constraints, does not separate into single-item subproblems because the product demands are substitutable.

Inventory management problems with substitutable demand, deterministic parameters, and serial transitive substitution structures have been studied as assortment problems. Examples include the works of Sadowski [1959], Wolfson [1965], and Tryfos [1985]. These have been extended to include stochastic demands ([Pentico 1974]) and two dimensional "square grid" substitution structures ([Pentico 1988]). Martel [1977] presented a problem with stochastic yield and stochastic demand but solved the problem by replacing the random yield variables with their expectations. The problem we pose is a generalization of the assortment problem.

Other problems with demand substitution include those investigated by McGillivray and Silver [1978], and Parlar and Goyal [1984]. They assumed that the production yields are known but only a fixed fraction of customers will accept substitute products. Product demands in these problems are stochastic.

The similarity among the problems above is that stochastic variables are on the demand side. Problems with randomness on the supply side are inherently more difficult and have been studied less. Yano and Lee [1989] reviewed lot-sizing problems with random yields. They revealed that the research in this area is concentrated on single-item, single-period, and uncapacitated cases. Our problems belong to the class of supply-side stochastic problems. We have the added features of multiple products, product demand substitutability, co-production, capacity and service constraints, and process selection. Product demand substitutability does

make the problem easier by permitting separation of some constraints. We demonstrate this later in the chapter. However, most of the features mentioned make determining the optimal solution of the problem very hard.

3. Problem description and Product Substitution Structure

The firm, that inspired this research, produces diodes for a variety of applications. These diodes are made from wafers of silicon or gallium arsenide. Chips sawn from the wafers are made into diodes. Each wafer contains about 5,000 chips. Every chip from the same wafer has the same physical design. The diodes derived from a wafer have different electrical characteristics because of process and material variations. Hence the yield rates of the chips from each wafer for any one product may be uncertain. The yield for a single product is usually low. When a family of products is considered as the outcome of a wafer, the total yield rate may be close to one. For this reason, we coordinate the production and inventory management decisions for the products.

The factory studied has about 30 product families. The wafers, as in most wafer fabrication facilities, are processed in batches or lots. As many as 12 wafers may be processed in each lot. We define a process as a set of machine settings, handling procedures, and materials used. All wafers in a lot undergo the same process. Most products require at most one lot for each period. Under these conditions, it is reasonable to assume that the probability distribution of the yield rates of products are independent of the lot size. Thus, the yield of a product in a lot can be obtained by multiplying the yield rate of the product by the production lot size. The processes are fairly stable and we will assume that the joint probability distribution of yield rates for any process is stationary over time.

All the product families share a production facility. The production capacity of this facility is a function of the number of lots and the time taken to process each lot. The processing time of each lot is independent of the lot size but each lot is limited by the number of wafers it can contain. For each product family, the lot processing times, under any process, are the same. We will examine each product family independently and assume, in each period, that a given number of lots has been allocated to the family. The allocation is made by a higher level planner.

The delivery schedule for the diodes is established in the contractual agreements with customers. The requirements for products are therefore determined for a horizon of 4 to 5 months. Alterations to the requirements are usually small. Consequently, we can assume that the demands are dynamic and deterministic, and we require that these demands be satisfied from inventory 100α% of the time. The latter is driven by objectives set by management. Another reason for using service constraints is that it is difficult to evaluate the penalties of not meeting delivery schedules for the products made in this facility.

The firm is installing sophisticated automated test equipment and information systems. These will be used for testing the chips and storing the summary test information. The tests are non-destructive. Each test gives the required electrical characteristics of individual chips and not just pass/fail results. The plan is to use the test equipment to select, from each wafer, the chips meeting the specifications of the product that the wafer was targeted for. After enough has been "cherry-picked" from the wafer, the remaining chips are made available to the "next best" use. There are subtleties as to how the cherry-picking should be done, how much is enough, and what is the "next best" use.

We described in the preceding chapter how some product substitution structures can be represented by acyclic directed graphs, G(V,E). V is the set of vertices and E is the set of directed edges. Each vertex represents a product. A directed edge from vertex i to vertex j implies that product i may substitute product j. Associated with each product i is a stock item i. Each process produces items according to an item yield rates joint probability distribution. Items may be used to satisfy the demand of their associated products or the demand of products that their associated products can be downgraded to. Figure 1 shows the relationships among process, items, and products. For transitive substitutions, if product i can substitute product j and product j can substitute product k, then product i can substitute product k. With this property, G(V,E) describes completely the inter-product relationships. This result is presented in chapter 2. However, without transitivity of substitution, this is no longer true.

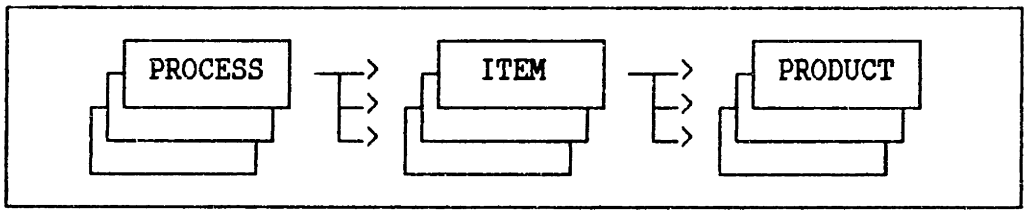


Figure 1. Process-Item-Product Chain Relationship.

Consider the two 3-products cases illustrated by figures 2a and 2b. In case a, the specifications of the products are 'nested' and in case b, the specifications overlap with each other without one containing another. The product substitution graph for case a has a serial structure. For case b, the interaction among the 3 product's specifications is in its worst possible configuration. It created 7 mutually exclusive subsets, labeled 1 through 7. Product 1 can accept all the units that meet the specifications of subsets 1, 4, 5, and 7. Similar statements can be made for products 2 and 3.

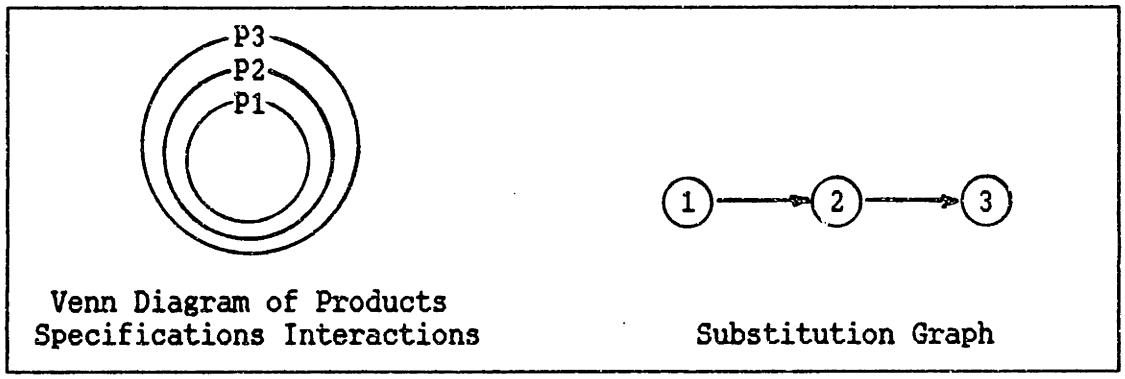


Figure 2a. Product Substitution Structure - Case a

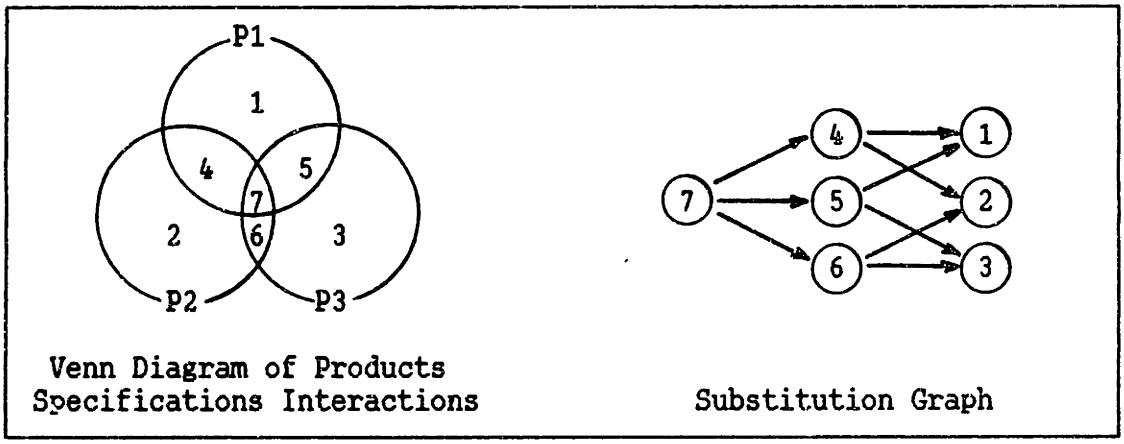


Figure 2b. Product Substitution Structure - Case b

From here on, we refer to the units that satisfy the specifications of subset i, as belonging to item i. The specifications for product i are now revised to be those of item i; originally, the specifications of product i comprised the union of the specifications of items or subsets that can be used by it. A new product is created for each item that have indices larger than those of actual products. In this example, only products 1, 2, and 3 are actual products; actual products have external demands. Henceforth, we refer to the actual products as real products and call the other products, pseudo-products. The term products is used to include both real and pseudo products. As in case a, under this representation, each product i has a corresponding item i and product demand substitution is transitive. Case b's substitution structure, for example, is a general acyclic directed graph. We can use the same approach

to construct transitive substitution structures for cases with any number of real products.

No assumptions are made about the nature of the specifications or that their limits be independent of each other. (See, for example, [Tang and Tang 1989] for cases of multi-characteristic product specifications where the limits are functions of more than one characteristic.) We have shown, without loss of generality, that substitution transitivity can always be made valid. Transitivity is achieved by creating mutually exclusive items. For n real products, we can have up to $2^n - 1$ items. In practice, the number of items is not so large because each product's specifications do not overlap with too many others. In the cases we have encountered, every product's specification overlap with no more than 2 others.

We now introduce some notation to represent subsets of products. The total number of items is n and the total number of real products is n_p. We define a(i) as the set of all products that can be directly downgraded to product i and b(i) as the set of all products that can be directly downgraded from product i. Products that can be directly downgraded from (to) product i have corresponding vertices one edge length away from vertex i in G(V,E) in (against) the direction of the directed edges. G(V,E) is the graph remaining after all redundant edges have been removed. An algorithm for removing redundant arcs was presented in [Bitran and Leong 1989]. A(i) is the set of all products that can be downgraded to product i and aggregate i, AU(i), is defined as equal to A(i)Ui. We define also B(i) to be the set of all products outside of AU(i) that can be directly downgraded from some $k \in AU(i)$. To ensure that units with alternative uses are not double-counted, we define the expanded aggregate AE(i) as equal to {i} if a(i) is empty, and $\{k:k \in AE(j), j \in a(i)\} \cup \{k:a(k) \in AE(j), j \in a(i)\}$, otherwise.

We show later how double-counting is eliminated. Crudely, AE(i) comprises the union of the aggregates that are of the same or higher hierarchical order, in the substitution structure, as product i. Figure 3 illustrates these definitions with an example.

Substitution Graph:

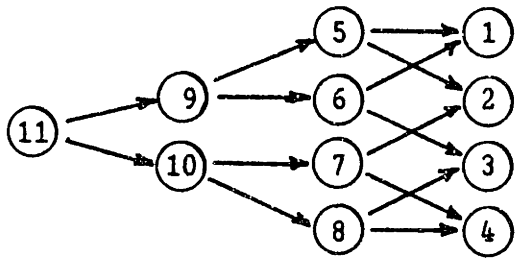


Table of Product Subsets:

i	a(i)	b(i)	A(i)	AU(i)	B(i)	AE(i)
1	5,6	-	5,6,9,11	1,5,6,9,11	2,3,10	1,..,11
2	5,7	-	5,7,9,10,11	2,5,7,9,10,11	1,4,6,8	1,..,11
3	6,8	-	6,8,9,10,11	3,6,8,9,10,11	4,5,7	1,..,11
4	7,8	-	7,8,10,11	4,7,8,10,11	2,3,9	1,..,11
5	9	1,2	9,11	5,9,11	1,2,6,10	5,..,11
6	9	1,3	9,11	6,9,11	1,3,5,10	5,..,11
7	10	2,4	10,11	7,10,11	2,4,8,9	5,..,11
8	10	3,4	10,11	8,10,11	3,4,7,9	5,..,11
9	11	5,6	11	9,11	5,6,10	9,10,11
10	11	7,8	11	10,11	7,8,9	9,10,11
11	-	9,10	-	11	9,10	11

Figure 3. Definitions of Subsets of Product - An Example

4. Model Formulation and Analytical Results

MODEL - PROBLEM FORMULATION

We define N_{st} as the production lot size for period t using process s. Associated with each process s, s=1,..,S, are yield rates q_{ist}, for items i=1,..,n_p and periods t=1,..,T. S is the total number of candidate processes and T is the length of the planning horizon. The demand for product i in period t is d_{it}. W_{ijt} is the amount to be downgraded from item i to j in period t. The unit holding and production costs are h and c respectively. F(x;y) is the cumulative density function of random variable

x evaluated at y. Prob(.) and E(.) are the probability and the expectation functions respectively. With these defined, we present below the stochastic linear programming formulation of the co-production problem.

(SPa)

$$Z_{SPa} = \text{Min } E(h\ \Sigma^{n}_{i=1}\Sigma^{T}_{t=1}\ J_{it}^{+} + c\ \Sigma^{T}_{t=1}\Sigma^{S}_{s=1}\ N_{st})$$

subject to

$$\text{Prob}(J_{it} \geq 0,\ i=1,..,n_p,\ t=1,..,T) \geq \alpha \qquad (1a)$$

$$\Sigma^{S}_{s=1}\ y_s \leq SU \qquad (2)$$

$$N_{st} \leq Q\ y_s, \qquad t=1,..,T,\ s=1,..,S \qquad (3)$$

$$N_{st},\ W_{ijt} \geq 0, \qquad (i,j) \in E,\ t=1,..,T,\ s=1,..,S \qquad (4)$$

$$y_s \geq 0,\ \text{integer} \qquad (5)$$

where SU is the number of lots allocated to the product family, Q is the maximum size of each lot, the net quantity of item i available at the end of period t

$$J_{it} = J_{i,t-1} + \Sigma^{S}_{s=1}\ q_{ist}N_{st} + \Sigma_{k\in a(i)}\ W_{kit} - \Sigma_{j\in b(i)}\ W_{ijt} - d_{it},$$

$$= \Sigma^{t}_{\tau=1}\ (\Sigma^{S}_{s=1}\ q_{is\tau}\ N_{s\tau} + \Sigma_{k\in a(i)}\ W_{ki\tau} - \Sigma_{j\in b(i)}\ W_{ij\tau} - d_{i\tau}),$$

$$i=1,..,n,\ t=1,..,T, \qquad (6)$$

and the inventory of item i at the end of period t

$$J_{it}^{+} = \max\{0,\ J_{it}\},\ i=1,..,n,\ t=1,..,T. \qquad (7)$$

The first constraint is the joint chance constraint on service and the next two are the capacity constraints. The service constraint means that the probability of any shortage is not more than 100(1-α)%. α, the probability target for meeting demand, is typically close to 1. In this formulation, we assumed that near-optimal solutions do not generate non-stationary accumulation of any item: the mean inventory level of any item will not increase with time. This requires that the processes available should be compatible with the demands. The alternative is to create a dummy product,

as a surrogate for scraps, that has an infinite demand and no service requirement, and incorporate it into the formulation.

Joint chance-constrained problems are difficult to solve because correlations in the variations of random variables make evaluation of the function hard. An alternative equivalent problem (SPb), focusing on aggregates, separates the service constraint into individual chance constraints.

(SPb)

$$Z_{SPb} = \text{Min } E(h \Sigma^{n}_{i=1}\Sigma^{T}_{t=1} J_{it}^{+} + c \Sigma^{T}_{t=1}\Sigma^{S}_{s=1} N_{st})$$

subject to

$$\text{Prob}(I_{it} \geq 0) \geq \alpha, \qquad i=1,..,n_p, \ t=1,..,T \tag{1b}$$

and constraints (2) to (5)

where

$$I_{it} = \Sigma_{k\epsilon AU(i)} J_{kt}$$

$$= I_{i-1,t} + \Sigma^{S}_{s=1} p_{ist} N_{st} - \Sigma_{j\epsilon B(i)} W_{ijt} - D_{it}$$

$$= \Sigma^{t}_{\tau=1} (\Sigma^{S}_{s=1} p_{is\tau} N_{s\tau} - \Sigma_{j\epsilon B(i)} W_{ij\tau} - D_{i\tau}),$$

$$i=1,..,n, \ t=1,..,T, \tag{8}$$

$$p_{ist} = \Sigma_{k\epsilon AU(i)} q_{kst}, \text{ and } D_{it} = \Sigma_{k\epsilon AU(i)} d_{kt}. \tag{9),(10}$$

Here I_{it}, p_{ist}, and D_{it} are aggregate variables for inventory, yield rate, and demand respectively. In the constraints of both formulations, we need only consider the items that correspond to real products since no external demands exist for pseudo-products.

Theorem 1: (SPa) and (SPb) are equivalent. ▪

For the sake of not disrupting the flow of this chapter, the proof of this and other theorems are presented in the appendix.

APPROXIMATIONS

We propose, in this section, deterministic approximations to problems

(SPa) and (SPb). These are linear programs amenable to any standard LP package. Before proceeding, further notation is introduced.

Notice, from figure 3, that some of the AE(.) sets are the same. We eliminate the redundant and append the distinct sets to AU(.) as AU(i), $i=n+1,..,n+n_e$ where n_e is the number of distinct AE(.). As a result, there are now $n+n_e$ aggregates. The distinct AE(.) sets can be constructed easily, from the substitution graph, using breadth-first search. We define $\phi_{is}(R)=F^{-1}(\Sigma^{R}_{r=1}p_{isr};1-\alpha)$ where $\phi_{is}(R)$ can be interpreted as the R periods $(1-\alpha)$ fractile for items good for product i. We let $\Omega_{is\tau k}$, $\tau=1,..,t$, and any i and s be defined as follows:

We construct t coefficients $\phi_{is}(1),..,\ \phi_{is}(1),\ (\phi_{is}(t)-(t-1).\phi_i(1))$. There are t possible permutations of these coefficients. We let $\Omega_{is\tau k}$, $\tau=1,..,t$, for each k, take on the values of the coefficients in the sequence presented by a permutation and set K(t)=t where $k \in \{1,..,K(t)\}$. (For example for t=2 and any i,s, $\Omega_{is11} = \phi_{is}(1)$, $\Omega_{is21} = (\phi_{is}(2) - \phi_{is}(1))$, $\Omega_{is12} = (\phi_{is}(2) - \phi_{is}(1))$, $\Omega_{is22} = \phi_{is}(1)$, and K(2)=2.)

An approximate stochastic linear program to (SPa) and (SPb) is

(SP1)

$$Z_{SP1} = \text{Min } E(h\ \Sigma^{n}_{i=1}\Sigma^{T}_{t=1}\ J_{it}^{+} + c\ \Sigma^{T}_{t=1}\Sigma^{S}_{s=1}\ N_{st})$$

subject to

$$M_{ist} - \Sigma^{t}_{\tau=1}\ \Omega_{is\tau k}\ N_{s\tau} \geq 0, \qquad i=1,..,n_p,\ t=1,..,T,\ s=1,..,S,\ k=1,..,K(t) \qquad (1c)$$

$$\Sigma^{S}_{s=1}\ M_{ist} - \Sigma^{t}_{\tau=1}\Sigma_{j\in B(i)}W_{ij\tau} \geq \Sigma^{t}_{\tau=1}D_{i\tau},\ i=1,..,n_p,\ t=1,..,T. \qquad (1d)$$

and constraints (2) to (5).

A deterministic approximation of (SP1) is

(DP1)

$$Z_{DP1} = \text{Min } h\ \Sigma^{n}_{i=1}\Sigma^{T}_{t=1}\Sigma^{t}_{\tau=1}[\Sigma^{S}_{s=1}E(q_{is\tau})N_{st} - d_{i\tau}] + c\ \Sigma^{T}_{t=1}\Sigma^{S}_{s=1}\ N_{st})$$

subject to constraints (1c), (1d) and (2) to (5).

The linear inequalities (1c) and (1d) are such that the extreme points they form are points at which selected rays from the origin intersect the lower boundary of (1b). The rays used are the axes of N_t, t=1,..,T and the ray in the center of the cones formed by these axes. The objective function of (DP1) is the same as Min ($h\ \Sigma^{n}_{i=1}\Sigma^{T}_{t=1}\ E(J_{it}) + c\ \Sigma^{T}_{t=1}\Sigma^{S}_{s=1}\ N_{st}$). It is made simpler by the fact that the unit inventory holding cost is a constant and that every "downgrading to" quantity has a corresponding "downgrading from" quantity. In (SP1) and (DP1), we assume in each period, each product's requirement is supplied mainly by one process. Hence we can consider the processes, in each period, independently without restricting the feasible region too much.

Theorem 2: If the feasible region of (SPa) is convex, any solution to (DP1) and (SP1) is feasible to (SPa). The same result is true for (SPb). ∎

The results of Monte-Carlo simulations, under the conditions of our test cases, indicate that the conditions of theorem 2 are reasonable for α close to 1. The common feasible region of (SPa) and (SPb) is, consequently, assumed convex for the rest of the chapter.

An equivalent of (DP1) is

(DP2)

$$Z_{DP2} = \text{Min } h\ \Sigma^{n}_{i=1}\Sigma^{T}_{t=1}\Sigma^{t}_{\tau=1}[\Sigma^{S}_{s=1}E(q_{is\tau})N_{st} - d_{i\tau}] + c\ \Sigma^{T}_{t=1}\Sigma^{S}_{s=1}\ N_{st})$$

subject to

$$M_{ist} - \Sigma^{t}_{\tau=1}\ \Omega_{is\tau k}\ N_{s\tau} \geq 0, i=1,..,n_p,n+1,..,n+n_e,\ t=1,..,T,$$
$$s=1,..,S,\ k=1,..,K(t) \qquad (1e)$$

$$\Sigma^{S}_{s=1}\ M_{ist} \geq \Sigma^{t}_{\tau=1}D_{i\tau}, \qquad i=1,..,n_p,n+1,..,n+n_e,\ t=1,..,T \qquad (1f)$$

and constraints (2) to (5).

Observe that problem (DP2) does not involve any downgrading variables. The number of variables in our problem has also been reduced. This is achieved by incorporating the concept of expanded aggregates AE(.), increasing the

number of constraints. From the computational viewpoint, it is easier to solve (DP1) than (DP2). (DP2) is, however, preferable because it is more intuitive; it can provide directions for constructing heuristics.

Theorem 3 [Upper bound on the relative error between the solutions of the stochastic and deterministic approximations.]: Let vector N^* be the optimal solution to the deterministic approximation (DP2) and vector W^* be such that (N^*,W^*) is a feasible solution in (SP1). The error of the value of the optimal solution to (DP2) relative to the value of the optimal solution to (SP1) is bounded above by $(ZU(N^*,W^*) - Z_{DP2})/Z_{DP2}$ where ZU(N,W) is the value of any feasible solution (N,W) to (SP1). ▪

The relative bound of theorem 3 indicates how well solutions of the deterministic approximation (DP2) will perform in practice when the stochastic approximation (SP1) is good. For α close to 1, the relative errors should be small. The computational experiments in chapter 2, using a similar approach on a simpler version of the problem in this chapter, suggest that the average of the relative error bound is around 3%.

The number of linear constraints, (1c) and (1d) (or (1e) and (1f)), to approximate the service constraints 1(a), is $O(n_p ST^2)$. The approximation may be refined by, as a result of introducing more rays, enlarging the set of inequalities. In fact, the original problem is reproduced if an infinite number of inequalities is used. Though the stochastic approximation can be made exact to the original problem, we do not attempt it for two reasons. The resulting program, firstly, will be very large and the computation time will be excessive. The second, more important, reason is that (SPa) and (SPb) are static problems. They do not take directly into account that later period decisions can be adapted to the state of the problem as it evolves. For the problem where decisions can be made sequentially, we

provide, below, a lower bound on the value of its optimal solution. This will be used as a benchmark for evaluating the heuristics. The theorem is derived by assuming the decision-maker has perfect control over the process.

Theorem 4 [Lower bound on the value of the optimal solution]: Let $qu = \text{Max}_s \{E[\Sigma^n_{i=1} q_{is.}]\}$ and $DU = E[\Sigma^n_{i=1} d_{i.}]$. The lower bound on the value of the optimal solution for the dynamic problem, Z_{LB} equals c (DU/qu). (Note that $d_{it}=0$ for $i=n_p+1,..,n$.) •

The notion of substitution among products can be extended to depict inventory transfer across periods. Holding an inventory so that the items may be used in later periods is in essence downgrading over time. Substitution over time has a serial structure and is transitive. We can expand the graphical representation of substitution relationships to include the time element. We re-label products such that every product i period t pair corresponds to a "product" (i,t). The new substitution graph has vertices (i,t) for "products" and "edges" ((i,t),(j,τ)) if product i in period t can substitute product j in period $\tau \geq t$. The graph can be reduced by the algorithm in chapter 2. After reduction, in any period, the edges between any two product should be as before and, across periods, there should only be edges between a product and itself over consecutive periods (i.e. to represent (i,t)->(i,t+1)). In this new graph, there will be nT "products", n_pT "real products", and up to (n_e+T) "expanded aggregates". Differing from the original problem, "downgrading" is no longer always free; "downgrading" across periods has holding cost.

5. Heuristics

All the heuristics are initialized by selecting a "best process" for each real product. We define the best process as the process that gives the

largest expected yield rate for the product. This reduces the number of eligible processes down, from about the number of items, to the number of real products. Eleven heuristics were tested and we report five significant ones. The first heuristic, P1NE, is the one being practiced by the facility studied; this is a common approach in industry. P1NE stocks inventory by products and considers demand one period only at a time. The production sizes are obtained by dividing the net demand by the expected yield rate. These decisions are made for each product independently and there is no inventory sharing. P1NF is the same as P1NE in all respect except that the production sizes are obtained using the process' $(1-\alpha)$ fractile rates. The fractile adjustment ensures that the service performance target is never exceeded.

Managers of the facility recognize that units not used by one product can be put to alternative uses. It has been proposed that units in excess of one product's demand should be systematically allocated to another product. P1NE can be modified to do this. The proposed heuristic, P1BE, "cherry-picks" enough good units to meet the immediate demand of the target product and allocates the remainders to the "next best" use. The "next best" product is the product with the highest expected yield rate for the target product's process. The product with the second highest expected yield becomes the "next best" if the product with the highest expected yield rate is the target product.

P1CF is another heuristic. As in P1NF, the fractile rate is used to ensure that the service performance target is satisfied. Unlike P1BE, all units that meet the specification of the target product are retained and stocked for that product. Only units that do not meet the target product's specifications are given away. This permit the build up of safety stocks, as was intended by the fractile correction. This heuristic shares most of

P1NF's features but has inventory sharing. The production decisions for P1BE and P1CF are also made independently for each product; they do not anticipate the possible fall-outs from other products' production.

Our final heuristic, M4DF, draws from the structural insight of the deterministic approximation (DP2) and the product substitution graph. For each product i, there is an item i, among the items product i can use, that has the lowest potential of being used by other products. In fact, sometimes item i can be used by product i only. Other items that can be delivered as product i, can also be used by other products. In this way, to satisfy product i's demand, item i should be preferred over other items. Generalizing, downgrading should be considered backwards along directed paths in G(V,E).

M4DF has a four period planning horizon and fractile-adjusted production sizes. It evaluates production sizing decisions for the products in descending order of their net demand and downgrades as mentioned in the preceding paragraph. Production, for each period, is limited by the capacity given. A sketch of the heuristic is as follows:

1. Compute the net demand of each product assuming each product, independently, has first claim on all items.
2. Take the product with the largest net demand and compute the production size, using the best process, for this product. If the production size is positive, set the inventory of the items usable for the target product to zero. Assuming the yield rates are at their fractile levels, update the inventory of the other items. If the computed production size is not positive, set it to zero and satisfy the target product's demand from the items, searching backwards in G(V,E). Now, set the net demand for the product, for both cases, to zero. Repeat step 1 until all products have been considered. The

production size of the processes thus computed will be referred to as the first period's estimates.

3. Steps 1 and 2 are repeated, considering this time the first two periods together. Compute the production sizes needed in the second period to meet demand, at the required service level, for periods 1 and 2 using inequalities (1e) and (1f). The starting inventory should be at their original levels and the first period's production is as estimated before. The resulting "estimated" second period production sizes are ranked in descending order of their sizes. In that order, the production sizes, of each process, are rounded off to the size of the nearest number of lots until the total number of lots used reaches the limit allowed. (By rounding off, we mean to bring it to the nearest integer. For example, 0.51 of a lot rounds off to 1 and 0.49 rounds off to 0.) The other process' production are set to zero. If the limit is not reached, the remaining lots are assigned, again in descending order of the size of the estimates, to the remaining processes that have positive estimated production size. These are the ones with production sizes greater than zero but less than one half of a lot. This trimming step ensures that capacity is never exceeded and the high demand products are satisfied first. A check-back step is then performed to increment the first period's production, with the second period's production as trimmed, to improve the second period's service level as much as possible. This is particularly for those products not produced in the second period because of limited capacity in that period. The revised first period's production sizes are now the first period's estimates.

4. The procedure in step 3 is repeated, adding one new period at a time. Each time, the production sizes of the period just added are

estimated with the first periods' as estimated and all others as trimmed. The estimates for this latest period's production are, as before, rounded off to the nearest number of lots and trimmed to meet capacity limits. Again, a check-back is used to revise the first period's estimates.

5. When all the periods in the horizon have been evaluated, the first periods production levels are tallied, in descending order of size, for the number of lots needed until the maximum number of lots is reached. The production sizes are rounded-up and not rounded-off as before. The remaining processes, even if they have positive estimated production size, are not activated.

M4DF is a fairly sophisticated yet simple heuristic. It coordinates production of the products within and across periods in the planning horizon. We propose M4DF as the heuristic to use for solving the co-production problem and will compare its performance against those of the other heuristics.

6. Computational Results and Comments

The heuristics were tested on a total of 270 cases. Details of these are given in the appendix. The simulations run for 500 periods, or approximately 42 years when each period corresponds to a month. In our test cases, we have assumed that the distributions and fluctuations of variables are random (uniformly distributed). The fractiles for the random variables and their convolutions were obtained through Monte-Carlo simulations. In practice, sample data may be used to estimate the multi-variate distributional form and the parameters for the random variables. Monte-Carlo simulations may be used, one-time off-line, to generate the fractiles when their close form expressions are not

available. Alternatively, distribution-free methods similar to those used in [Allen, Braswell, and Rao 1979] or parametric approximation methods as proposed in [Pinter 1989] can be used.

A summary of the results is given in figure 4. Figure 4 reports only 240 of the 270 cases. The 30 unreported cases have lot size of 30 and allocations of 2 or 3 lots per period. These cases are "infeasible": always violates the capacity and service constraints. In the simulations, we track the cost, capacity, and service performance. Cost performance is reported in measures relative to the lower bound. ZR(heuristic), the value of the objective function, for each heuristic, relative to the lower bound, equals $Z_{heuristic}/Z_{LB}$. The top table in figure 4 shows the values of ZR(.) for each test case operating under each heuristic. The bottom table lists the number of cases, out of 15 cases, that each heuristic violates the service and capacity constraints. Figures 5 to 9 graphed the results to ease inference.

[INSERT FIGURE 4 HERE]

As shown in figure 5, P1NE and P1NF cost about the same. The detail results showed that P1NF satisfies service performance target whereas P1NE always violate it. Similarly, we found that the results of P1BE, in all the cases and for all measures of performance, always dominate those of P1NE. Therefore, P1NF and P1BE improve upon P1NE with almost no additional penalty. These are easy improvements and can be implemented quickly. P1BE is lower in cost than P1NF but violates service limits. P1CF was designed to rectify this weakness in P1BE. However, as shown in figure 5, P1CF cost outcome do not have a nice relationship to that of P1NE. P1CF's results, evidently, can be very bad. An explanation for this is that P1CF, because

it does not coordinate the production of the products, tends to build up too much inventory. Hence, P1CF ensures the operation meet the service target but, in doing so, it may incur large additional cost. When the processes used produce very few "by-products", P1CF is no worse than P1NF.

M4DF has features that overcome the short-comings of the other heuristics. First, it complies with the capacity limits; none of the heuristics mentioned make any provision for capacity. However, because of this, under very tight capacity M4DF will violate service limits. For the test problems, this did not occur even when the capacity is as low as 1.2 times of average total demand. M4DF, therefore, provides excellent service, keeps within the capacity limits, and does so at costs lower than that of all the other heuristics. This is illustrated by figures 5 to 8.

[INSERT FIGURES 5 TO 8 HERE]

Figure 9 shows that the costs for M4DF fall monotonically as capacity is increased. For the practically uncapacitated cases, the cost of operating under M4DF is about only 14% above the lower bound. With tight capacity, M4DF's cost is about 30% above the lower bound. For P1NE and P1NF, regardless of capacity, the results are 78% and 81% respectively. The actual costs for P1NE and P1NF should actually be much higher since, in the simulations, these heuristics violate capacity with no penalty. The relative error of M4DF is, therefore, substantially smaller; it is less than half those of P1NE and P1NF.

[INSERT FIGURE 9 HERE]

The bound given in theorem 4 is actually a very poor lower bound. So the results of M4DF can be very close to optimal. We saw from heuristic M4DF that a trade-off between service (or capacity) performance and cost is not always necessary. We must highlight that M4DF stocks inventory in more categories. Additional costs are incurred to maintain the larger inventory system. These costs are not explicitly included in our model. Finally, we observed that the gains from using better coordinated methods are small if the yield rates of the best process for each product are very close to one. In such cases, performance of the production system is good even when the products' production are planned independently.

7. Summary

We showed that, under a simple transformation, problems made complicated by intertwined product specifications can be reduced to structures with transitive substitutions. Restructuring and representing the relationships as acyclic directed graphs, provide a congenial framework for coordinating the decisions for the products.

M4DF emulates a deterministic approximation and demonstrates to be a very good heuristic. For the cases tested, it costs only between 14 to 33% more than the lower bound. Since the lower bound is quite loose, such deviation is probably small. The dynamic process selection approach using the best process for each product and evaluating products in descending order of their net demand size seems adequate. It is also fairly easy to implement.

In conclusion, this chapter attempted to bring new insights to the concepts of quality and flexibility in manufacturing. As more manufacturers go for narrower segments of markets, the need to understand how the proliferation of very specific product offerings can impact production and

allocation decisions becomes greater. This chapter demonstrates an example of this type of analyses. The chapter discusses a problem in a semiconductor manufacturing context. Extensions can be made for applications in other manufacturing and service operations. This will be the topic for the next chapter.

REFERENCES

ALLEN, F.M., R. BRASWELL, and P. RAO 1974. "Distribution-free Approximations for Chance-Constraints," Op. Res. 22, 610-621.

BITRAN, G.R. and S. DASU 1989. "Order Policies in an Environment of Stochastic Yields and Substitutable Demands," Working paper #3019-89-MS, Sloan School of Management.

BITRAN, G.R. and H.H. YANASSE 1984. "Deterministic approximations to stochastic production problems," Op. Res. 32, 999-1018.

DEUERMEYER, B.L. and W.P. PIERSKALLA 1978. "A By-Product Production System with an Alternative," Mgmt. Sci. 24, 1373-1383.

MARTEL, A. 1977. "A Probabilitic Assortment Problem," INFOR 15, 196-203.

McGILLIVRAY, A.R. and E.A. SILVER 1978. "Some concepts for inventory control under substitutable demand," INFOR 16, 47-63.

PARLAR, M. and S.K. GOYAL 1984. "Optimal ordering decisions for two substitutable products with stochastic demands," Opsearch 21(1).

PENTICO, D.W. 1974. "The assortment problem with probabilistic demands," Mgmt. Sci. 21, 286-290.

PENTICO, D.W. 1988. "The discrete two dimensional assortment problem," Oper. Res. 36, 324-332.

PINTER, J. 1989. "Deterministic Approximations of Probability Inequalities," Methods and Models of Operations Research 33, 219-239.

SADOWSKI, W. 1959. "A few remarks on the assortment problem," Mgmt. Sci. 6, 13-24.

SILVER, E.A. and R. PETERSON 1985. Decision Systems for Inventory Management and Production Planning, 2nd ed., J. Wiley.

TANG, K. and J. TANG 1989. "Design of Product Specifications for Multi-characteristic Inspection," Mgmt. Sci. 35, 743-756.

TOPKIS, D.M. 1968. "Optimal ordering and Rationing Policies in a non-stationary Dynamic Inventory model with n Demand classes," Mgmt. Sci. 15, 160-176.

TRYFOS, P. 1985. "On the optimal choice of sizes," Oper. Res. 33, 678-684.

VEINOTT, A.F. 1965. "Optimal Policy in a Dynamic, Single product, non-stationary Inventory model with several demand classes," Oper. Res. 13, 761-778.

WOLFSON, M.L. 1965. "Selecting the best length to stock," Oper. Res. 13, 570-585.

YANO, C.A. and H.L. LEE 1989. "Lot-Sizing with Random Yields: A Review," Technical Report 89-16, Dept. of Industrial and Operations Engineering, U. of Michigan.

APPENDIX

Proof of Theorem 1:

No external demands for items $i=n_p+1,..,n$; $N_{st}\geq 0$, $s=1,..,S$, $t=1,..,T$; $q_{it}\geq 0$, $i=1,..,n$, $t=1,..,T$ and $\text{Prob}(J_{it}\geq 0,\ i=1,..,n_p,\ t=1,..,T) \geq \alpha$ $\Rightarrow \text{Prob}(J_{it}\geq 0,\ i=1,..,n,\ t=1,..,T) \geq \alpha$.

Similarly, no external demands for items $i=n_p+1,..,n$; $N_{st}\geq 0$, $s=1,..,S$, $t=1,..,T$; $p_{it}\geq 0$, $i=1,..,n$, $t=1,..,T$ and $\text{Prob}(I_{it}\geq 0) \geq \alpha$ for $i=1,..,n_p$, $t=1,..,T$ $\Rightarrow \text{Prob}(I_{it}\geq 0) \geq \alpha$, $i=1,..,n$, $t=1,..,T$.

($\Rightarrow$) For any i and t, $\text{Prob}(J_{it}\geq 0,\ i=1,..,n,\ t=1,..,T) \geq \alpha \Rightarrow \text{Prob}(J_{kt}\geq 0,\ k\epsilon A(i)Ui) \geq \alpha$. By definition $I_{it} = \Sigma_{k\epsilon AU(i)}J_{kt}$. Hence $\text{Prob}(I_{it}\geq 0) \geq \alpha$ for any i and t.

($\Leftarrow$) For any i and t, we know that $\text{Prob}(I_{kt}\geq 0) \geq \alpha$ for $k \epsilon A(i)$. For those $k \epsilon A(i)$ such that $\text{Prob}(I_{kt}\geq 0) > \alpha$, we can downgrade some of their units to product i till $\text{Prob}(I_{kt}\geq 0) = \alpha$. Hence we can make $\text{Prob}(I_{kt}\geq 0) = \alpha$ for all $k \epsilon A(i)$ without changing the objective value. But $\text{Prob}(I_{it}\geq 0)$ can only increase with downgrading from above. $\text{Prob}(I_{it}\geq 0) \geq \alpha$, $\text{Prob}(I_{kt}\geq 0) = \alpha$ for all $k \epsilon A(i)$, and $I_{it} = \Sigma_{k\epsilon A(i)}Ui J_{kt}$ implies $\text{Prob}(J_{it}\geq 0) \geq \alpha$.

Therefore, (SPa) is equivalent to (SPb). ▪

Proof of Theorem 2:

The feasible regions of (SP1) and (DP1) are polyhedrons with extreme points on the surface on the lower boundary of the feasible region of (SPa). Since the feasible region of (SPa) is convex, any solution to (DP1) and (SP1) is also feasible to (SPa). Same argument goes for (SPb). ▪

Proof of Theorem 3:

Consider the problems,

(DP3+)

$Z_{SPc+} = \text{Min } (h\ \Sigma^{n}_{i=1}\Sigma^{T}_{t=1}\ (E(J_{it}))^{+} + c\ \Sigma^{T}_{t=1}\Sigma^{S}_{s=1}\ N_{st})$

subject to constraints (1e), (1f) and (2) to (5).

(DP3)

$Z_{SPc} = \text{Min } (h\ \Sigma^{n}_{i=1}\Sigma^{T}_{t=1}\ (E(J_{it})) + c\ \Sigma^{T}_{t=1}\Sigma^{S}_{s=1}\ N_{st})$

subject to constraints (1e), (1f) and (2) to (5).

(DP3) is the same as (DP3+) except for $(.)^{+}$ in the objective of (DP3+). Therefore, $Z_{DP3} \leq Z_{DP3+}$. (DP3+) is the same as (SPa) except that the expectation is taken before taking $(.)^{+}$. By convexity of J_{it}^{+} and Jensen's inequality, $Z_{DP3+} \leq Z_{SP1}$. Hence, $Z_{DP3} \leq Z_{DP3+} \leq Z_{SP1}$.

For every "downgrading to" variable, there is a "downgrading from" variable. By this, we can show that the objective functions in (DP2) and (DP3) are the same. The feasible regions of (DP2) and (DP3) are identical. Consequently, (DP2) and (DP3) are equivalents and $Z_{DP2} \leq Z_{SP1}$.

N^{*} optimal to (DP2) corresponds to a feasible solution (N^{*},W^{*}) in (SP1). Therefore, $ZU(N^{*},W^{*}) \geq Z_{SP1} \geq Z_{DP2}$.

The relative error, $RE \equiv (Z_{SP1} - Z_{DP2})/Z_{SP1} \leq (ZU(N^{*},W^{*}) - Z_{DP2})/Z_{DP2}$. ∎

Proof of Theorem 4:

We let the inventory in each period be zero. This assumes that the capacity is unlimited, there is one process (the one with the best overall expected yield rate) and the decision-maker gets, from the process, the items in the relative proportions desired. Hence, in the long run, the average total cost is the average production cost. ∎

Test Cases

There are 3 groups of 90 test cases, making a total of 270. Each test case has 4 products, indexed 1 through 4, and 11 items as shown in figure 3. An additional item, item 12, is created to represent rejects. The total demand of the products is assumed to be uniformly distributed between 750 and 1250 units, with a mean of 1000 and a range of 500. Three set of weights - (1,1,1,1),(1,5,15,40), and (20,1,20,1) - are used to generate product demands for the three groups. The 4 weights in each set are the demand weights for the 4 products. In each group, the demand is determined by assigning the total demand to the products in the relative proportions of 4 randomly generated numbers weighted by the demand weights. The demands are given for four periods into the future.

The 90 cases, in each group, are divided into 5 equal sub-groups. Each sub-group shares a set of 10 candidate processes. The process capability is given as a set of 12 numbers; one for each item. These numbers or weights are randomly generated and permanently assigned to the process. For each period, the yield rates under each of these processes is generated as follows. A random number is generated for each item. The weighted proportion of these random numbers using the weights according to its process capability is used as the yield rate of outcome of the process. We test the problems with 2, 3, 4, 5, 8, and 16 lots allowed per period and lot size of 30, 60, and 90. Six number of lot levels, 3 lot size levels and 5 sub-groups give a total of 90 cases in each group.

For these test cases, the $(1-\alpha)$ fractiles are generated by Monte-Carlo simulations. The service performance target, α, is set at 0.95 and the unit production and holding costs are 8 and 1 respectively. The starting inventory position of all items, before initialization, are zero.

The system is initialized with 50 simulated periods of use. The simulation runs for 500 periods.

VALUE OF OBJECTIVE FUNCTION RELATIVE TO THE LOWER BOUND, ZR(.)

CASE	P1HE	P1HF	P1BE	P1CF	MHOF with lot size=30 MAX. NO. OF LOTS 4	5	8	16	MHOF with lot size=60 MAX. NO. OF LOTS 2	3	4	5	8	16	MHOF with lot size=90 MAX. NO. OF LOTS 2	3	4	5	8	16
T01	1.04	1.07	1.62	1.38	1.05	1.05	1.24	1.13	1.34	1.30	1.23	1.19	1.13	1.12	1.30	1.21	1.15	1.13	1.11	1.11
T02	1.75	1.78	1.57	3.21	1.06	1.00	1.26	1.10	1.04	1.02	1.24	1.22	1.10	1.10	1.00	1.22	1.19	1.10	1.17	1.17
T03	1.70	1.73	1.51	1.02	1.04	1.01	1.19	1.12	1.01	1.24	1.17	1.15	1.12	1.12	1.22	1.16	1.13	1.12	1.11	1.11
T04	1.02	1.06	1.70	2.00	1.06	1.07	1.25	1.17	1.04	1.04	1.20	1.21	1.17	1.17	1.01	1.22	1.19	1.17	1.16	1.17
T05	1.70	1.01	1.60	3.93	1.26	1.25	1.24	1.15	1.20	1.26	1.20	1.20	1.15	1.14	1.26	1.20	1.16	1.14	1.14	1.14
T06	1.00	1.04	1.67	1.44	1.05	1.06	1.25	1.15	1.03	1.02	1.26	1.22	1.15	1.14	1.01	1.22	1.17	1.13	1.13	1.13
T07	1.75	1.79	1.50	2.44	1.09	1.09	1.26	1.19	1.06	1.03	1.25	1.23	1.19	1.10	1.01	1.25	1.20	1.19	1.10	1.10
T08	1.67	1.71	1.50	1.30	1.34	1.02	1.21	1.13	1.00	1.25	1.10	1.16	1.12	1.12	1.22	1.16	1.13	1.12	1.12	1.11
T09	1.02	1.06	1.71	2.77	1.07	1.09	1.20	1.10	1.04	1.04	1.27	1.23	1.17	1.17	1.02	1.24	1.20	1.17	1.17	1.17
T10	1.77	1.01	1.50	3.44	1.20	1.22	1.23	1.14	1.20	1.24	1.22	1.19	1.13	1.13	1.25	1.21	1.15	1.13	1.12	1.12
T11	1.00	1.04	1.60	1.45	1.06	1.03	1.23	1.12	1.04	1.20	1.22	1.10	1.11	1.11	1.20	1.21	1.14	1.12	1.10	1.10
T12	1.74	1.77	1.56	3.00	1.00	1.09	1.26	1.10	1.06	1.03	1.25	1.21	1.10	1.17	1.02	1.22	1.20	1.10	1.17	1.17
T13	1.67	1.70	1.49	1.30	1.34	1.02	1.20	1.12	1.01	1.25	1.19	1.15	1.12	1.12	1.20	1.17	1.13	1.12	1.12	1.12
T14	1.79	1.03	1.67	2.01	1.30	1.00	1.20	1.17	1.05	1.04	1.24	1.22	1.17	1.10	1.31	1.22	1.19	1.17	1.16	1.10
T15	1.76	1.00	1.59	5.40	1.24	1.20	1.22	1.10	1.25	1.23	1.21	1.17	1.12	1.11	1.24	1.19	1.15	1.12	1.11	1.11
AVE	1.70	1.01	1.60	2.49	1.03	1.03	1.24	1.15	1.02	1.29	1.23	1.20	1.15	1.14	1.20	1.21	1.17	1.15	1.14	1.14

NUMBER OF CASES NOT MEETING SERVICE AND CAPACITY CONSTRAINTS--OUT OF 15 TEST CASES

		Lot size=30 MAX. NO. OF LOTS 4	5	8	16	Lot size=60 MAX. NO. OF LOTS 2	3	4	5	8	16	Lot size=90 MAX. NO. OF LOTS 2	3	4	5	8	16
P1HE	: SERVICE	15	15	15	15	15	15	15	15	15	15	15	15	15	15	15	15
	: CAPACITY	15	15	15	0	15	15	15	15	0	0	15	15	15	13	0	0
P1HF	: SERVICE	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	: CAPACITY	15	15	15	0	15	15	15	15	2	0	15	15	15	15	0	0
P1BE	: SERVICE	15	15	15	15	15	15	15	15	15	15	15	15	15	15	15	15
	: CAPACITY	15	15	15	0	15	15	15	15	0	0	15	15	15	0	0	0
P1CF	: SERVICE	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	: CAPACITY	15	15	15	0	15	15	15	15	0	0	15	15	15	0	0	0
MHOF	: SERVICE	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	: CAPACITY	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Figure 4. Summary of Simulation Results

Figure 5. Cost Comparison among Heuristics.

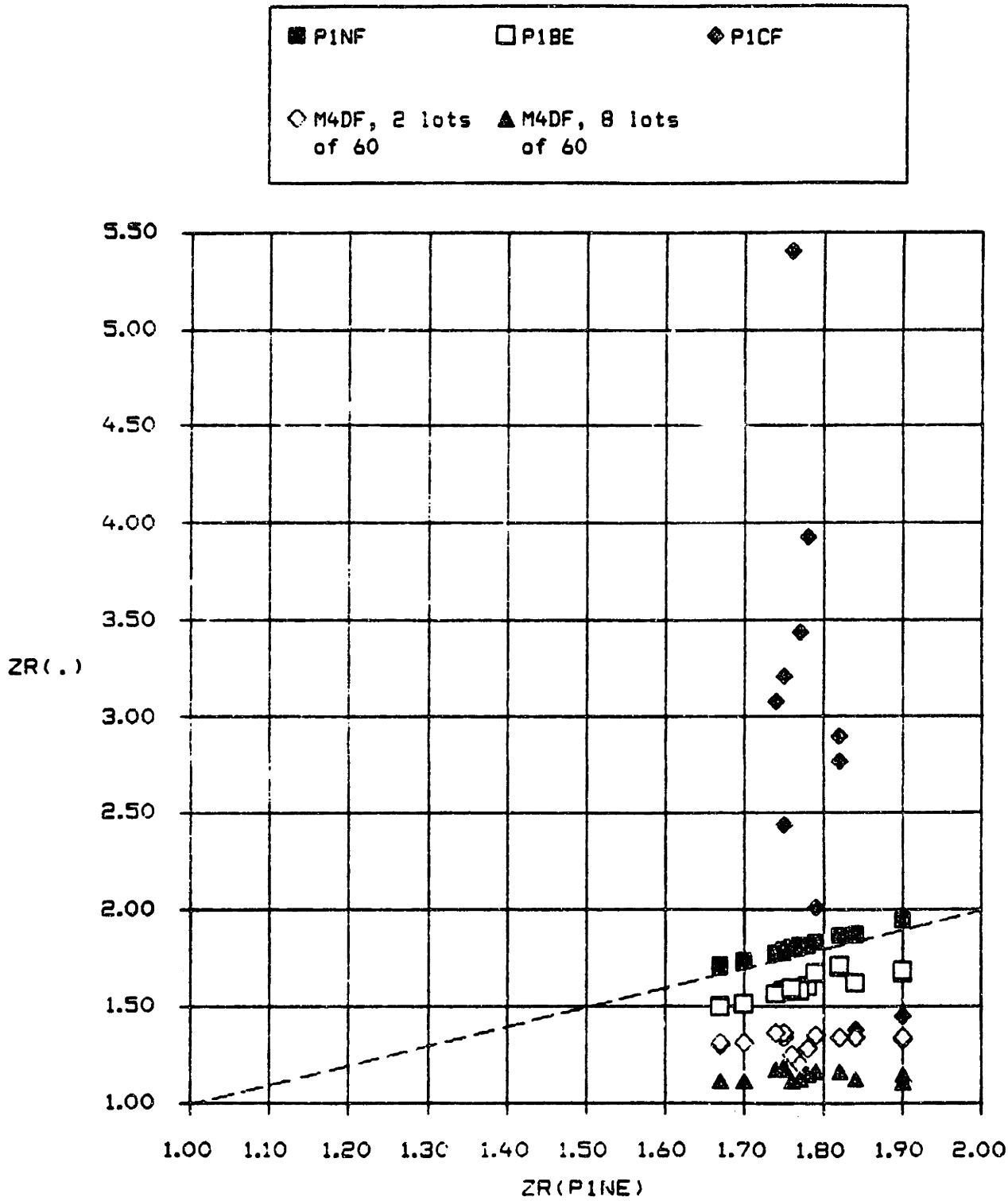


Figure 6. M4DF against P1NF -- lot size of 30.

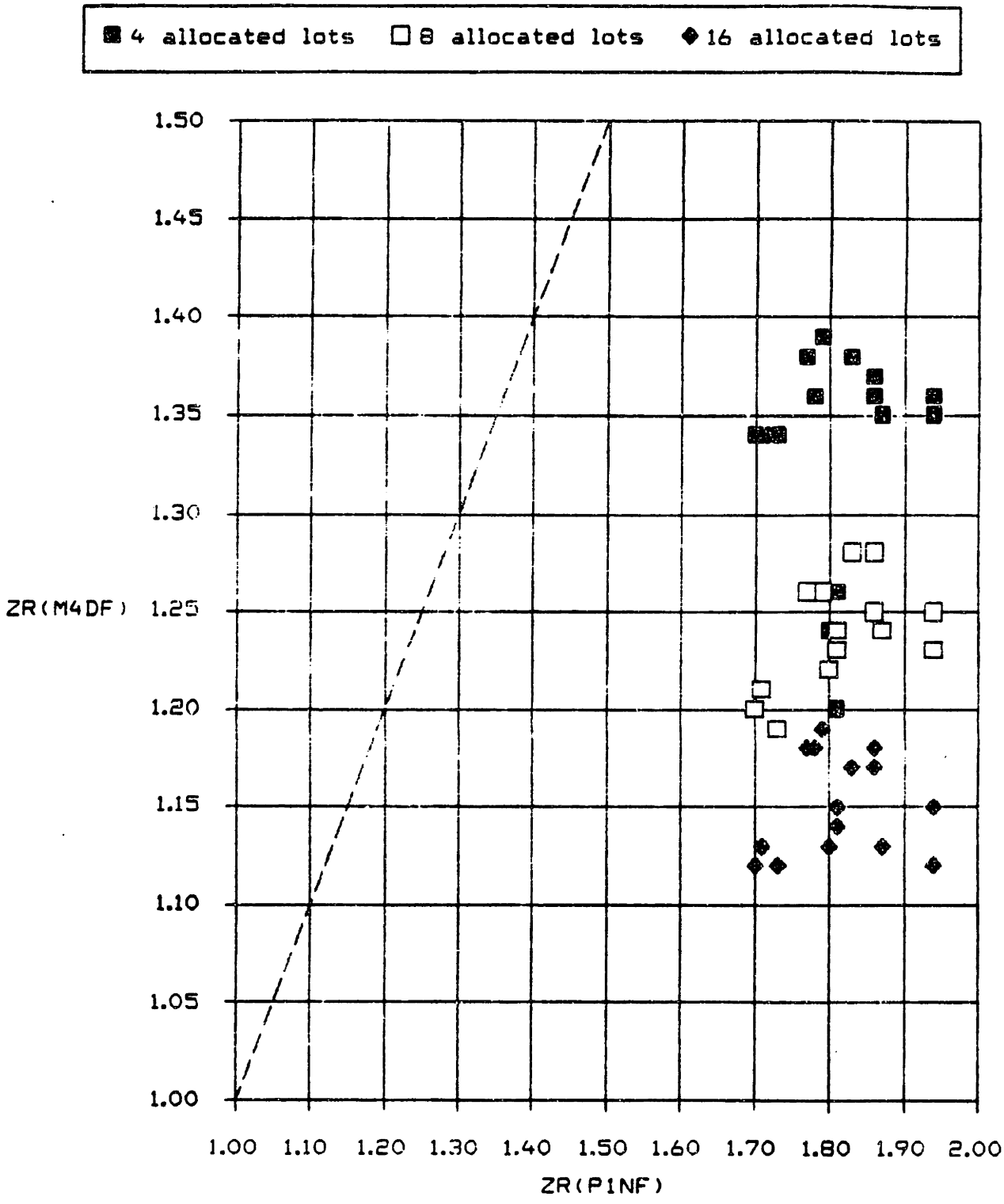


Figure 7. M4DF against P1NF -- lot size of 60.

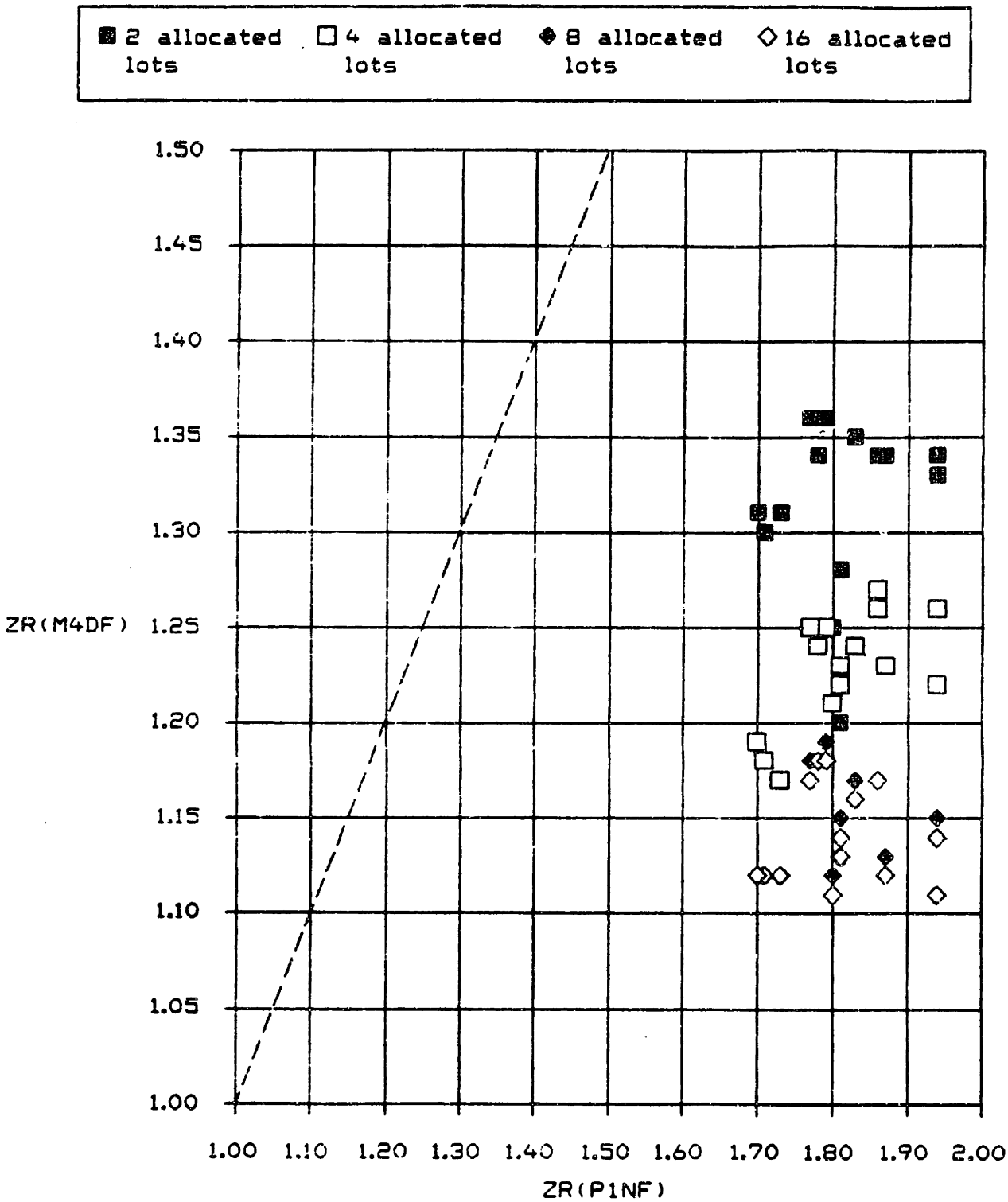


Figure 8. M4DF against P1NF -- lot size of 90.

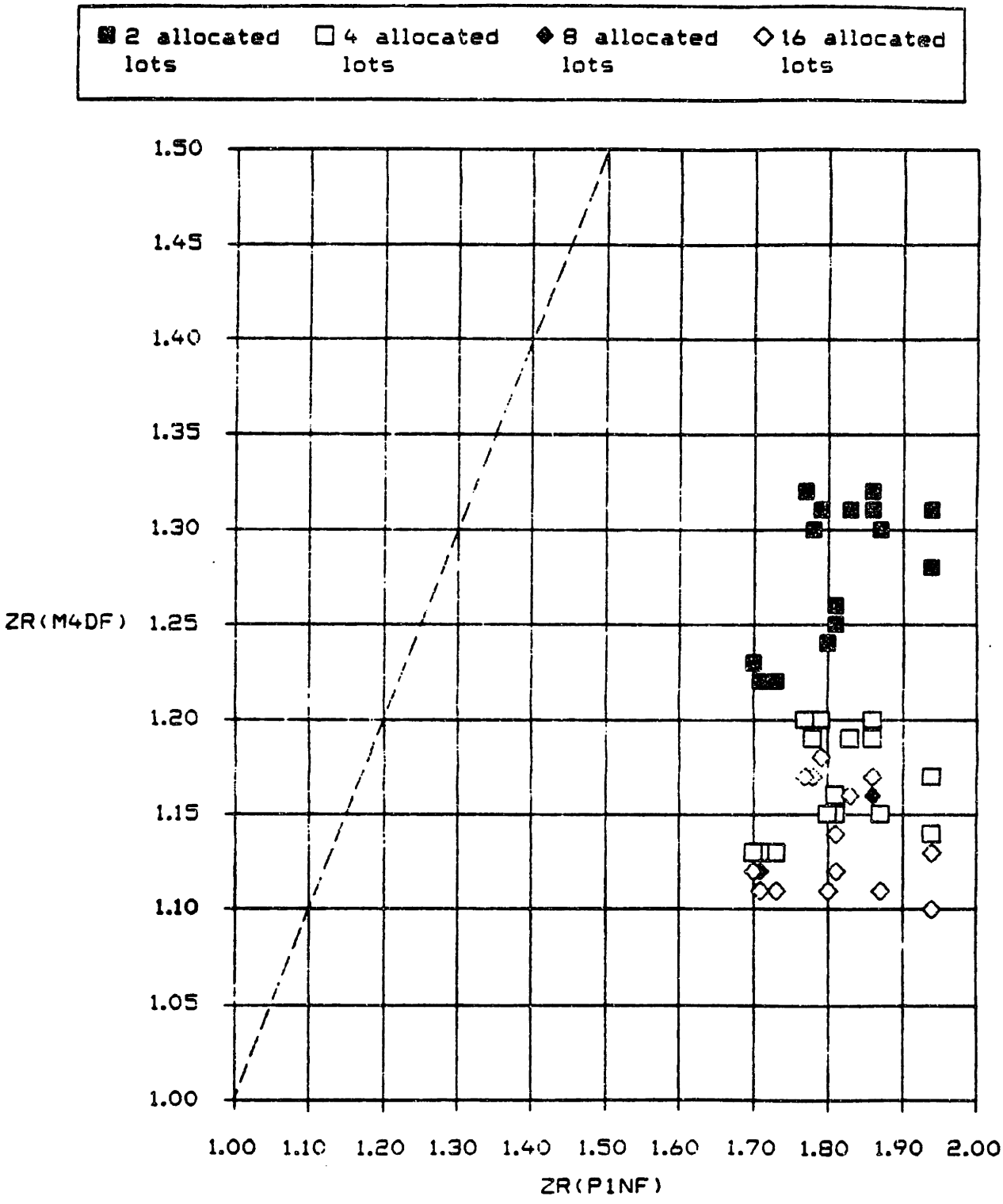
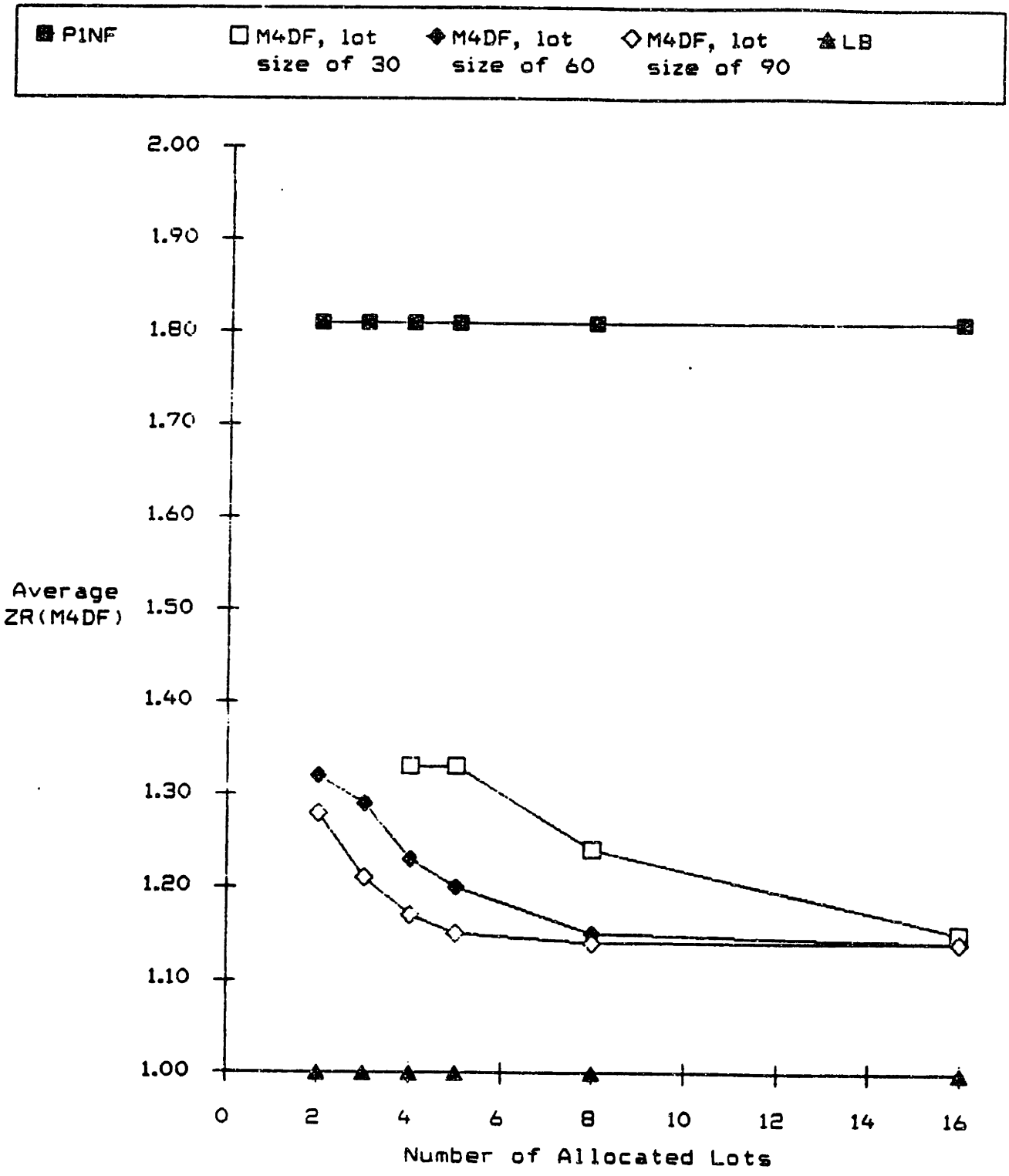


Figure 9. Average ZR(M4DF) at different capacities.



CHAPTER 4

HOTEL SALES AND RESERVATIONS PLANNING

1. Introduction

Hotels take room reservations from a few months to one day in advance. Prospective guests can cancel their reservations anytime before the day the rooms are required; cancellations are made with no penalty. Prospective guests, without informing the hotel, may even fail to show up for their reservations. The number of cancellations and no-shows can be highly variable. Though expected no-show rate is around 15%, Rothstein [1974] quoted estimates from hotel executives that no-show rates in excess of 25% are common, indicating the problem's magnitude and difficulty.

Other major sources of room demand are stayover and walk-in. Occasionally, trips must be taken on short notice, forcing the traveler to seek accommodations as a walk-in, a prospective guest with no reservation. Even when a guest makes and honors a reservation, the estimated length of stay may be inaccurate. A business executive who planned a three day visit, for example, may take four days to settle her affairs, thus making it necessary to extend the room occupation. Conversely, she may finish in two days, permitting early departure. Therefore, these room demands are also random.

Even though the major sources of demand are random, some types of demand can be controlled. Reservation demand is controlled by limiting the number of reservations to accept. Stayovers cannot exceed the number of rooms currently occupied. This, in turn, depends on the number of reservations previously accepted. Some hotels, in policy, honor all

requests for stay extension. But most hotels, depending on capacity available, may or may not extend a stay beyond what was scheduled. (When this general practice is resisted, hotels will usually back-off to avoid unnecessary negative publicity. Occurrences like these are rare and may be neglected.) As such, the hotels have some control over stay extensions. Similarly, walk-in demands can be selectively rejected when there is insufficient capacity. Premature departures, on the other hand, cannot be directly controlled. So as to get enough time to adapt, most operators set rules on the amount of pre-checkout notice their guests must give.

In sharp contrast to the consumer's right of cancellation without penalty, a hotel, on the other hand, is obligated to live up to its reservation commitments. To remain competitive and profitable, it is prudent that hotels plan how they run the reservation operation. We propose that they plan the booking of reservations, to complement the other demands. We aim towards maximizing expected profit subject to service constraints for meeting the demand from booked reservations. We believe that this is a novel formulation for the hotel problem.

The problem is related to the production planning problem with stochastic yields. The number of reservations to accept corresponds to the production lot size and no-shows correspond to rejects. Reservations accepted and guests present are equivalents of stocking items. These stocks "perish" when there are cancellations or premature departures. Unlike manufacturing of products, services such as hotel room "rentals" cannot be produced ahead of time and stocked in anticipation of seasonal demand. Hence capacity not utilized is lost forever; pre-emptive production is not possible. Furthermore, since there is no backordering, demands not met are also lost forever. From this comparison, we see that the hotel reservation

problem is richer and more interesting than the production planning problem.

This chapter is organized as follows. We review in section 2 the literature related to the hotel reservation problem. Section 3 describes the problem that we intend to solve. In section 4, we formulate the problem as linear programs and present the main results. Additional comments and extensions are given in section 5. We end the chapter with a summary and conclusions.

2. Literature Review

Rothstein [1974] claimed that he found no published model directed specifically to the hotel problem and provided one. His model is an extension of the airline overbooking problem examined previously by Rothstein [1968, 1971a, 1971b]. He used the Markovian sequential decision process to generate booking policies for hotels with one room-type and single-day stays. This problem differs from the airline problem by allowing double occupancy--more than one guest per room.

Ladany [1976] extended Rothstein's airline work to provide a hotel model where there are two room-types: single and double rooms. Stay durations are still limited to single-days only. The author claimed that the model may be extended for many room-types and multiple-day stays. The state space for this dynamic program will be huge. One study that explicitly model stays of more than one period is [Kinberg, Rao, and Sudit 1980]. In this model, there are two categories of demand: package (subscription) and spot. The model determines how the fixed resource capacity should be allocated to the two demand categories. Subscriptions are sold with price discounts, but are paid in advance; the trade-off is between degree of demand uncertainty and expected total revenue. The

problem is fundamentally different from ours in that tickets sold are paid; no-shows do not create problems. Glover et al. [1982] and Pfeifer [1989] studied how airlines should allocate capacity to different fare classes. Again, these problems do not consider cancellations and show uncertainties.

Liberman and Yechiali [1978] allow hotels to cancel confirmed reservations or acquire additional reservations. Both are done with penalties to the hotel. With identical rooms and focusing on a single target date, they showed that the optimal policy consists of 3 regions demarcated by 2 threshold numbers. The regions are where the options--(a) accept all new requests and acquire additional reservations, (b) do nothing, and (c) cancel some confirmed reservations--are appropriate. This model is essentially an extension of the well-known newsvendor problem. Buying and selling of reservations may be viewed as an indirect approach of incorporating the multiple room-types feature in a one room-type model.

William's [1977] model is the most complete, considering practically all the major sources of demand. However, his model assumes that there is only one type of room. He evaluated the problem on three separate criteria: expected cost, expected underbook and number of walks, and expected occupancy rate and number of walks. Walks are people who have made reservations but cannot check-in because of room shortages; they walk away dissatisfied. The most interesting outcome from William's work is a set of histograms and smoothed approximations constructed from data obtained from two hotels. He showed that reservations, scheduled stayovers, and unscheduled stayovers show-rates can be approximated by Beta distributions; and walk-ins follow the Gamma distribution. Scheduled stayover show-rate is one minus premature departure rate.

Even though the works mentioned studied service operations, they and most others do not incorporate explicit measures on service performance.

Exceptions include the work by Thompson [1961], Taylor [1962], Shlifer and Vardi [1975], and Jennings [1981]. Thompson, who initiated the approach, studied control issues in airline reservations. He provides feasible solutions to the problem with two seat-classes that has constraints on the risk of exceeding capacity. No cost parameter or objective function is present in this problem and the problems in the other papers mentioned in this paragraph. Single flight-leg problems, in these papers, are similar to one period hotel problems; multiple flight-legs problems are similar to multiple periods problems.

In general, the airline problem has a lot of features in common with the hotel problem. The interested reader should refer to [Rothstein 1985] for a review of that problem. Other related problems include hospital admissions and bed allocations ([Kao and Tung 1981]), clinic appointment systems ([Rising, Baron and Averill 1973]), and car or equipment rentals ([Tainter 1964] and [Whisler 1967]).

In this chapter, we draw upon the parallel between the hotel problem and the manufacturing problem solved in chapter 2. The problem considered in that chapter has random production yield and substitutable product demand. Unlike previous hotel reservation studies, the formulation we provide for our hotel problem has multiple periods, room-types, and guest-classes. New features addressed, not found in the manufacturing problem, include perishability of inventory, no pre-emptive production, and multiple recourse opportunities. Also, in manufacturing terminology, the related production model backorders when there are shortages whereas hotels has lost-sales.

We alluded to the first two features in the introduction. We now mention briefly what multiple recourse opportunities mean. Reservations, made in advance, may be cancelled by the guest before the required day.

However, as long as that day is still in the future, additional reservations can be accepted, to make up for those cancelled. So the hotel model, unlike the manufacturing analogue we mentioned, has multiple opportunities to respond to a demand--room-type for a certain day.

3. Problem description

Hotel rooms are frequently classified into types: suite, deluxe, and standard rooms, to suit different lifestyles and budgets. When a prospective guest with reservation, arriving in good time, finds no available room in the hotel, an oversale is said to have taken place. Oversale occurs because hotels sometimes overbook reservations to keep occupancy levels high. When oversale of a particular room-type occurs, hotel operators can choose between turning away the prospective guest or giving her, at no additional cost, a better room. The first option must be mitigated with an offer of alternative accommodation--at a competing hotel--and freebies, for example, a free dinner at the hotel's restaurant. In addition to loss of revenue and extra costs, the fear of goodwill loss makes hotel management desire to see this happen as rarely as possible. "Downgrading" a room, on the other hand, adds a contribution to profit though smaller than what it is potentially capable of. Nevertheless, the downgraded room may have remained vacant and contributed nothing.

We classify hotel rooms into ordered types $s \in \{1,..,m\}$ where 1 is the most luxurious and m the least. A room from each room-type may be offered at more than one rate. The rates are different because of the nature of occupancy (single/double/with children), discounts, commissions, and costs of extra promotion. We also classify the market into ordered classes $i \in \{1,..,n\}$. Now, we let $a(s)$, $s=1,..,m$, be the indices of classes

such that guests in classes a(s),..,a(s+1)-1 request for room-type s where $1=a(1) < a(2),.., < a(m) \leq n$.

Class i guests pay c_i per room for each night of occupancy. The guest-classes for the same room-type are labeled in descending order of the rates charged; guests for room-type s may be charged any of the rates c_i, $i \in \{a(s),..,a(s+1)-1\}$. The highest rate for each room-type is often referred to as the rack rate for that room-type. When guests of more luxurious rooms always pay more for their rooms than guests of less luxurious rooms, by the labeling convention, $c_i \geq c_j$ if $i < j$. This need not always be true.

The reader should note that classes are not necessarily defined according to rates alone: market segments that compete for the same room-type and pay the same rates may be classified as different classes. The classes defined, however, must not be discriminatory against individuals and, at the time of receiving a reservation request, the hotel operator should be able to distinguish which class the request belongs to. For example, Shlifer and Vardi [1975] mention that, because of the significant differences in their cancellation and show behaviors, reservations from different geographical origins have been classified into different classes.

Figure 1 demonstrates, with an example, the relationship among the room-types and guest-classes. Each vertex represents a guest-class. A directed edge leading from vertex i to vertex j represents the possibility that a room allocated to class i can be offered to class j. By virtue of the labeling order of room-types and guest-classes, there is a directed edge from every class i to i+1. That is, a class i guest paying class i rate, but offered a room that is acceptable to class j guests will not be dissatisfied if $j < i$.

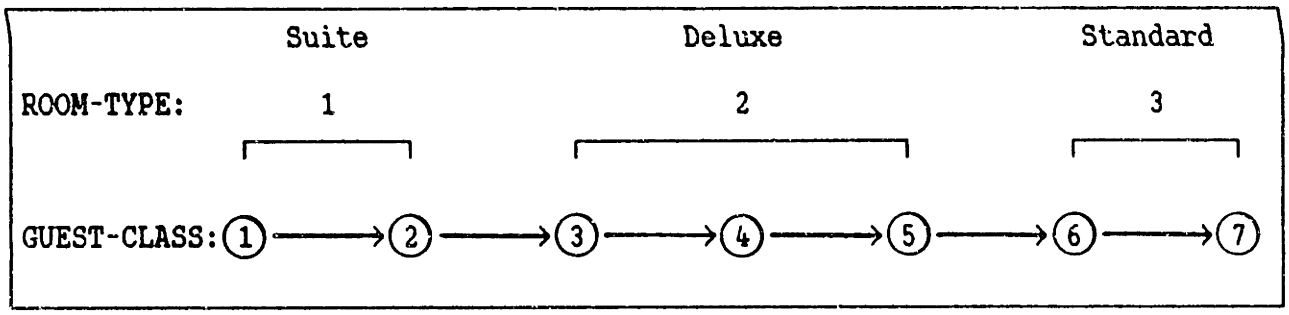


Figure 1. Room-types and Guest-classes--An example

Set aside for prospective class i guests are N_{it}, i=1,..,n t=1,..,T, number of reservations for period t at the start of period t. T is the length of the planning horizon. The number of class i guests who will show up in period t is q_{it} N_{it} where q_{it} is the class i reservation show-rate for period t. Show-rate q_{it} ϵ [0,1] is a continuous random variable. The yield or show size is given as a product of the show-rate and the size of the reservation. This assumes that the yield rate distribution is independent of the reservation size. Liberman and Yechiali [1978] made the same assumption and William [1977] provided empirical evidence that this assumption is reasonable. We will also assume that reservation and show sizes are sufficiently large that the requirement for the decision variables to be integers may be relaxed.

4. Model

The purpose of our model is to assist in the planning for the optimal level of reservations and in appropriating the hotel's capacity to market segments. These decisions support both the sales and operations functions. We propose to solve the problem in two stages: (a) reservations planning, and (b) walk-in control. We end this section with additional guidelines for managing sales and setting room rates.

RESERVATIONS PLANNING

For a given horizon, we first work toward getting the optimal

reservation target levels. Operators are then "authorized" to accept reservations up to these levels. The target levels suggest how the capacity of the hotel should be allocated to the guest-classes; the reservations targets are attainable only when there is sufficient demand. We do not consider any specific assignment of rooms to the reservations since the reservations may be cancelled or may not show. By default, the capacity remaining is for walk-in guests. Walk-in guests, usually charged rack rates, may have a significant portion of the rooms set aside for them. Unlike the airline reservations problem, hotels do not always need to overbook reservations because the walk-in demand, fetching high returns, can be substantial.

Prospective guests make reservations for period t, to stay for one or more days. For a planned stay of s periods, we record the reservation as separate individual reservations for periods t,..,t+s-1. The justification for doing this comes from the empirical evidence in [William 1977]. In figures 1 and 2 of his paper, William fitted Beta distributions to the two sets of show-rate data and showed that the fits are excellent. The mean and coefficient of variation of the fitted reservations show-rate distribution are 0.83 and 0.083 respectively. The corresponding statistics for scheduled stayover show-rate are 0.86 and 0.083. Hence, scheduled stayovers and reservations have probability distributions that are practically identical. It is also reasonable to assume that the two show-rates are statistically independent.

For the rest of this chapter, we refer to the combined show-rate distribution of reservations and scheduled stayovers as simply the show-rate distribution. When a guest with multiple-days booking did not show on the first day of the intended stay or cuts short the scheduled stay, the bookings for the remaining days are considered cancelled. Booked

reservations, being commitments, are given the highest priority when conflict arises. The second priority goes to walk-ins. Stay extensions have the lowest priority: hotels are not bound to satisfy stay extension requests. No service performance limits are set for meeting stay extension requests; stay extension inquiries will be treated as if they are new reservation requests.

We define MU_{st} as the number of type s rooms available in period t and M_{it} to be the number of rooms initially allocated to guest-class i for period t. We set $M_{it} = MU_{st}$ for i = a(s) and $M_{it} = 0$ otherwise, s=1,..,m, t=1,..,T. In this way, we allocate the rooms to the highest guest-class possible and we make the rooms available indirectly to the other classes through W_{it}, the number of rooms from those allocated to class i to downgrade to class i+1 during period t.

We define NS_{it} as the random variable for the demand of guest-class i reservations in period t and YS_{it} as the number of class i prospective guests that will walk into the hotel during period t without reservations. NS_{it} and YS_{it}, i=1,..,n and t=1,..,T, have finite mean and variance, and are random variables in $[0,\infty)$. Figure 2 shows the sources of demand by class i prospects for rooms in period t.

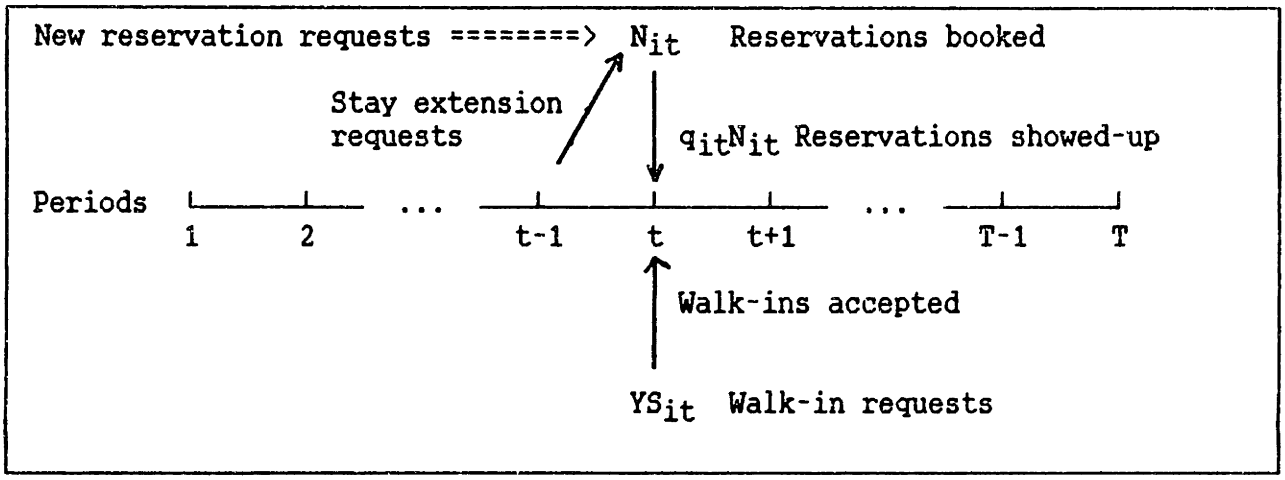


Figure 2. Demand by class i prospects for rooms in period t.

For simplicity of presentation, we let $A(s) = \{a(s),..,a(s+1)-1\}$ and $AU(s) = \{1,..,a(s+1)-1\}$, $s=1,..,m$. $A(s)$ is the set of all guest classes for type s rooms and $AU(s)$ is the set of all guest-classes that can be offered type s rooms. We present, below, a stochastic linear programming formulation of the reservations problem.

(SPa)

$$Z_{SPa} = \text{Max } \{E_q[E_{NS}[\Sigma^n_{i=1}\Sigma^T_{t=1} c_{it} (M_{it} + W_{i-1,t} - W_{it} - J_{it}^+)] \quad (1)$$

$$+ \text{Max } \Sigma^n_{i=1}\Sigma^T_{t=1} c_{it} E_{YS}[Y_{it}]] \quad (2)$$

subject to

$$Y_{it} \le YS_{it}, \quad i=1,..,n, \ t=1,..,T, \quad (3)$$

$$0 \le Y_{it} \le J_{it}^+, \ i=1,..,n, \ t=1,..,T \ \} \quad (4)$$

subject to

$$N_{it} \le NS_{it}, \quad i=1,..,n, \ t=1,..,T, \quad (5)$$

$$\text{Prob}(J_{it} \ge 0, \ i=1,..,n) \ge \alpha, \quad t=1,..,T, \quad (6)$$

$$W_{it}, N_{it} \ge 0, \quad i=1,..,n, \ t=1,..,T, \quad (7)$$

where $E_x[.]$ is the expectation function over the random vector x; and q, NS, YS are the vectors of random variables q_{it}, NS_{it}, YS_{it} respectively.

Also, $W_{0t} = 0$, $t=1,..,T$,

$$J_{it} = M_{it} + W_{i-1,t} - W_{it} - q_{it} N_{it}, \quad i=1,..,n, \ t=1,..,T, \quad (8)$$

and

$$J_{it}^+ = \text{Max}(0, J_{it}), \quad i=1,..,n, \ t=1,..,T. \quad (9)$$

In the reservations problem (SPa), we optimize the total expected revenue, by allocating rooms among reservation and walk-in prospects. This is subjected to service constraints to ensure that revenues are not increased by making reservation promises that the hotel cannot usually keep.

when $J_{it} \geq 0$, and $c_{it}(M_{it} + W_{i-1,t} - W_{it})$ when $J_{it} < 0$; with some algebraic manipulation and taking expectation results in (1). We call (2) to (4), the sub-problem in (SPa), (RPa): (RPa) is the walk-in recourse problem. Constraint (3) states that walk-ins accepted cannot exceed walk-in requests. Constraint (4) ensures the capacity of the hotel is not exceeded and walk-ins cannot take negative values. Constraint (5) makes certain that the reservations booked cannot exceed reservations requested. The service constraint, (6), guarantees that oversale occurs with less than $100(1-\alpha)\%$ probability. $\alpha \in [0,1]$ is the service performance target for booked reservations and, set according to management policy, should be close to 1. Constraint (7) are non-negativity constraints for the decision variables in the main problem. The other equations, self-explanatory, are introduced to simplify the presentation.

Notice that constraints (3) and (5) have stochastic right-hand-side terms that must not be violated. Therefore, other than the trivial zero reservations and zero walk-ins solution, there is no other feasible solution to (SPa). As such problem (SPa) has no meaningful solutions; we will reformulate the problem slightly. Before proceeding further, we present some important results of the reservations planning problem.

Theorem 1 [Time period separation]: Problems (SPa) separates into T one-period problems. ▪

This theorem suggests that reservations planning can be executed by focussing on one period at a time. In view of this, the results of earlier papers that focus on single-period problems may be valid. Therefore, by theorem 1, we drop the period index, t, and focus on a particular period of interest--referred to, from here on, as the target period. All subsequent reference to equations will be made as if index t does not exist.

interest--referred to, from here on, as the target period. All subsequent reference to equations will be made as if index t does not exist.

Theorem 2 [Joint chance constraint separation]: Constraint (6) is equivalent to $\text{Prob}(\Sigma^{i}_{j=1}\, q_j\, N_j \leq \Sigma^{i}_{j=1}\, M_j) \geq \alpha$, $i=1,..,n$.

Proof of theorem 2: Constraint (6) => $\text{Prob}(\Sigma^{i}_{j=1}\, J_j \geq 0) \geq \alpha$, $i=1,..,n$. By equation (8), $\text{Prob}(\Sigma^{i}_{j=1}\, (M_j - W_j - q_j\, N_j) \geq 0) \geq \alpha$, $i=1,..,n$. By the non-negativity constraint (7), $W_i \geq 0$. Hence the result. The converse is true using similar arguments and downgrading when necessary. ∎

Theorem 2 provides an alternative way of expressing the service constraint. The resulting separation of the original joint chance constraint into individual chance constraints makes the problem more tractable.

We now reformulate the problem by incorporating constraints (3) and (5) into the objective function but first we introduce more notation. We define f(x;y) and F(x;y) to be the value of the probability density and the cumulative density functions respectively for any random variable x evaluated at y. We let $YS_i(y) = YS_i$ for $YS_i \leq y$ and $YS_i(y) = y$ otherwise, $i=1,..,n$. Using this new variable, we can insert (3) into the $E_{YS}[Y_i]$ term in the objective function of (RPa) to give (RPb).

(RPb)

$$Z_{RPb} = \text{Max}\ \Sigma^{n}_{i=1}\, c_i\, E_{YS}[YS_i(Y_i)] \qquad (2a)$$

subject to

$$0 \leq Y_i \leq J_i^{+}, \qquad i=1,..,n. \qquad (4a)$$

Theorem 3: $E_{YS}[YS_i(y)]$, $i=1,..,n$, is non-decreasing, concave in y.

Proof of theorem 3: The first derivative of $E[YS_i(y)] = 1 - F(YS_i;y) \geq 0$. Also, the second derivative of $E[YS_i(y)] = - f(YS_i;y) \leq 0$. Therefore, $E[YS_i(y)]$, $s=1,..,m$, is non-decreasing and concave in y. ∎

Using the results of the theorems above, we re-write (SPa) to give (SPb).

(SPb)

$$Z_{SPb} = \text{Max } E_q[E_{NS}[\Sigma^n_{i=1}\, c_i(M_i + W_{i-1} - W_i - J_i^+)] + \Sigma^n_{i=1}\, c_i\, E[YS_i(J_i^+)]] \quad (1a)$$

subject to

$$N_i \le NS_i, \qquad i=1,..,n, \quad (5a)$$

$$\text{Prob}(\Sigma^i_{j=1}\, q_j\, N_j \le \Sigma^i_{j=1}\, M_j) \ge \alpha, \qquad i=1,..,n, \quad (6a)$$

$$W_i,\ N_i \ge 0, \qquad i=1,..,n. \quad (7)$$

We mentioned, in the first paragraph of section 3, that oversales are usually mitigated with offers of alternative accommodations. Up to now, we have not included the cost of oversales into the problem. This cost, except for the more explicit components, is usually quite difficult to quantify. We will assume, from here on, that the cost of oversale for each guest-class is its room rate. This is an attempt to capture as much of the quantifiable costs as possible. Of course, we already have constraints to ensure that the service goals are met--an indirect way of acknowledging the more esoteric costs. The resulting program differs from (SPb) by the absence of the $(.)^+$ function for the second term in the objective function. We also repeat the approach used to reformulate (RPa) to incorporate constraint (5) into the objective function. We present the final formulation as (SPc).

(SPc)

$$Z_{SPc} = \text{Max } E_q[E_{NS}[\Sigma^n_{i=1}\, c_i(M_i + W_{i-1} - W_i - J_i(N_i)^+)] + \Sigma^n_{i=1}\, c_i\, E[YS_i(J_i)]] \quad (1b)$$

subject to

$$\text{Prob}(\Sigma^i_{j=1}\, q_j\, N_j \le \Sigma^i_{j=1}\, M_j) \ge \alpha, \qquad i=1,..,n, \quad (6a)$$

$$W_i,\ N_i \ge 0, \qquad i=1,..,n, \quad (7a)$$

where $NS_i(y) = NS_i$ for $NS_i \le y$ and $NS_i(y) = y$ otherwise, $i=1,..,n$, (10)

(SPc)

$$Z_{SPc}=\text{Max } E_q[E_{NS}[\Sigma^n_{i=1}\ c_i(M_i+W_{i-1}-W_i-J_i(N_i)^+)] + \Sigma^n_{i=1}\ c_i\ E[YS_i(J_i)]] \quad (1b)$$

subject to

$$\text{Prob}(\Sigma^i_{j=1}\ q_j\ N_j \le \Sigma^i_{j=1}\ M_j) \ge \alpha, \quad i=1,..,n, \quad (6a)$$

$$W_i,\ N_i \ge 0, \quad i=1,..,n, \quad (7a)$$

where $NS_i(y) = NS_i$ for $NS_i \le y$ and $NS_i(y) = y$ otherwise, $i=1,..,n$, (10)

$$J_i(y) = M_i + W_{i-1} - W_i - q_i\ NS_i(y), \quad i=1,..,n, \quad (11)$$

and

$$J_i(y)^+ = \text{Max}(0,\ J_i(y)), \quad i=1,..,n. \quad (12)$$

APPROXIMATIONS

We propose two approximations: stochastic and deterministic. Each approximation leads us progressively towards a tractable problem. (SP), the stochastic program, approximates (SPc) by linearizing the feasible region of (SPc). The user chooses how accurate the approximation should be. At the expense of doing an infinite amount of work, (SP) becomes (SPc). In our experience, a simple approximation like the one we are presenting has small relative errors. The deterministic approximation (DP1) approximates (SP) by simplifying the objective function. (DP1) is formulated so that an upper bound on the relative error between (SP) and (DP1) can be obtained. This result is presented in theorem 6. Lastly, we linearize the separable convex program (DP1) into a deterministic linear program, (DP2).

To construct (SP), we replace, for each i, the service constraint (6a) by a set of linear constraints. This set of linear constraints is uniformly tighter than the original constraint it replaces; any solution feasible to the set of linear constraints is also feasible to the original constraint. The detail of this inner-linearization approach is discussed in chapter 5. The approach is as follows: We define $\phi(a_1,..,a_n) = F^{-1}(\Sigma^n_{i=1}\ a_i$

(SP)

$$Z_{SP}=\text{Max } E_q[E_{NS}[\Sigma^n_{i=1}\ c_i\ (M_i+W_{i-1}-W_i-J_i(N_i)^+)] + \Sigma^n_{i=1}\ c_i\ E[YS_i(J_i)]] \quad (1b)$$

subject to

$$\Sigma^i_{j=1}\ \Omega_{jk}\ N_j \le \Sigma^i_{j=1}\ M_j,\quad k=1,..,K1(i),\ i\quad ..,n \quad (6b)$$

$$W_i,\ N_i \ge 0, \qquad i=1,..,n \quad (7a)$$

Theorem 5: $E_{NS}[NS_i(y)]$, i=1,..,n, is non-decreasing, concave in y.

Proof of theorem 5: Same as in theorem 3. ∎

For α sufficiently close to 1, by constraints (6), (6a), or (6b) and the presence of a recourse problem, the capacity allocation guarantees that oversale seldom happen: $J_i \ge 0$ most of the time. As an approximation, we will assume that $J_i \ge 0$ for all i. Next, we remove the outer most expectation function and take expectation of variable J_i.

(DP1)

$$Z_{DP1} = \text{Max } \Sigma^n_{i=1}\ c_i\ (E[q_i]\ E_{NS}[NS_i(N_i)] + E_{YS}[YS_i(E_q[J_i])]]) \quad (1c)$$

subject to

$$\Sigma^i_{j=1}\ \Omega_{jk}\ N_j \le \Sigma^i_{j=1}\ M_j,\quad k=1,..,K1(i),\ i=1,..,n \quad (6b)$$

$$W_i,\ N_i \ge 0, \qquad i=1,..,n \quad (7a)$$

Theorem 6 [Upper bound on the relative error between the value of the optimal solutions to (SP) and (DP1)]: Let vector N^* be the optimal solution to (DP1) and vector W^* be such that (N^*,W^*) is a feasible solution in (SP). The relative error between the values of the optimal solutions to (DP1) and (SP) is bounded from above by $(Z_{DP1} - ZU(N^*,W^*))/ZU(N^*,W^*)$ where $ZU(N^*,W^*)$ is the value of (N^*,W^*) in (SP).

Proof of theorem 5: We call upon the convex properties of functions $(.)^+$, and theorems 3 and 5 to apply Jensen's inequality. ∎

By theorems 3 and 5, using a standard approach in separable convex programming, we linearize the objective function: (a) the first term in

(1c) is replaced by $\Sigma^{n}_{i=1}\Sigma^{K2(i)}_{k=1}\, d_{ik}\, x_{ik}$ where d_{ik}, $d_{i1} > .. > d_{i,K2(i)}$, are new cost coefficients and x_{ik}, $0 \le x_{ik} \le xu_{ik}$, i=1,..,n, k=1,..,K2(i) are the new variables; (b) the second term in (1c) is replaced by $\Sigma^{n}_{i=1}\Sigma^{K3(i)}_{k=1}\, e_{ik}\, y_{ik}$ where e_{ik}, $e_{i1} > .. > e_{i,K3(i)}$, are new cost coefficients and y_{ik}, $0 \le y_{ik} \le yu_{ik}$, i=1,..,n, k=1,..,K3(i) are the new variables. Note that $N_i = \Sigma^{K2(i)}_{k=1}\, x_{ik}$ and each x_{ik} is contained in a given partition where the expected marginal return is approximately d_{ik}. Similarly, $E[J_i] = \Sigma^{K3(i)}_{k=1}\, y_{ik}$ and each y_{ik} is contained in a partition where the expected marginal return is approximately e_{ik}. $K_{2(i)}$ and $K_{3(i)}$ are the number of piecewise-linear segments used to approximate each of the corresponding functions. After making the approximations, we simplify and present the new problem as (DP2).

(DP2)

$$Z_{DP2} = \text{Max } \Sigma^{n}_{i=1} \left(\Sigma^{K2(i)}_{k=1}\, d_{ik}\, x_{ik} + \Sigma^{K3(i)}_{k=1}\, e_{ik}\, y_{ik}\right) \qquad (1d)$$

subject to

$$\Sigma^{i}_{j=1}\, \Omega_{jk}\, N_j \le \Sigma^{i}_{j=1}\, M_j, \qquad k=1,..,K1(i),\ i=1,..,n \qquad (6b)$$

$$\Sigma^{K2(i)}_{k=1}\, x_{ik} = N_i, \qquad i=1,..,n \qquad (13)$$

$$0 \le x_{ik} \le xu_{ik}, \qquad k=1,..,K2(i),\ i=1,..,n \qquad (14)$$

$$\Sigma^{i}_{j=1}\Sigma^{K3(j)}_{k=1}\, y_{jk} \le \Sigma^{i}_{j=1}\, (M_j - E[q_j]\, N_j), \qquad i=1,..,n \qquad (15)$$

$$0 \le y_{ik} \le yu_{ik}, \qquad k=1,..,K3(i),\ i=1,..,n \qquad (16)$$

$$N_i \ge 0, \qquad i=1,..,n \qquad (17)$$

In practice, hotels designate some capacity for walk-ins and then, basing on the remaining capacity, estimate how many reservations to accept. (DP2) does the same thing but achieve it with an analytical approach.

Given the reservation targets, the desired operational response is to control the external and stay-extension requests for reservations, by reacting to cancellations. This aim of the exercise is to have the

reservation levels, for each day at the start of that day, hit their respective targets. This is impossible when there are insufficient requests. Even when there are enough requests, it is difficult, using the approaches currently practiced, to attain these targets because the cancellations are random. The approaches in use usually accept reservations, for periods far into the future, up to some authorization level. The authorization level is usually given as a fixed percentage above available capacity. In reality because of cancellations, authorization levels, rather than flat over time, should be larger the further away the current period is from the target period.

Accepting early bookings increases the certainty of getting enough business. Examples of early booking sources are package tour operators and convention organizers. These early bookings tend to fetch lower rates and, therefore, hotels may refuse some of them in the hope of getting more lucrative business later. The demand from the later market segments may be very uncertain and hence the need to trade-off. To include this trade-off into our model, so as to give better authorization levels, we broaden the concept of show-rate.

Show-rate was defined in conjunction with the definition of N_i: it was defined as the fraction of reservations still 'alive' at the start of the target period that will show up by the end of that period. There are two time-points of reference here: an end point and a start point. The end point is the end of the target period and the start point is the point the reservation targets are set for. Since we are usually concerned about the reservation targets for the beginning of the current period, we will call the start time-point the current period.

The broader concept, the survival rate, introduced now, involves both the cancellation and the show characteristics of reservations. We say a

reservation survived if it has not been cancelled or failed to show. For a target period, the survival-rate, q_i, is the fraction of reservations that will survive from among the reservations that were "alive" now (at the current period) plus those to be accepted from now until the target period. With this amendment, the reservation targets obtained from the programs will be the authorization levels for the current period--and not, as previously defined, for the start of the target period. The earlier definition is a special case of this extended definition.

WALK-IN CONTROL

Walk-ins targets are not explicitly specified in the solution of our problem. In this sub-section, to assist in the control of walk-in demand, we present a decision rule. This rule helps hotel operators decide how to allocate rooms to the requests by different class of walk-ins and, in particular, suggests when rooms should be downgraded for walk-ins. Consider the problem (C1).

(C1)

$$Z_{C1}(Y_i,Y_j)$$

$$= c_i \int_0^{Y_i} y\, f(YS_i;y)dy + c_i\, Y_i \int_{Y_i}^{\infty} f(YS_i;y)dy$$

$$+ c_j \int_0^{Y_j} y\, f(YS_j;y)dy + c_j\, Y_j \int_{Y_j}^{\infty} f(YS_j;y)dy$$

$$+ \mu_i\, (LY_i - Y_i)$$

$$+ \mu_j\, (LY_j - Y_j)$$

where

$i < j$, $j=2,..,n$,

Y_i is the number of rooms to offer to class i walk-ins, $i=1,..,n$,

LY_i is the capacity available for class i walk-in, $i=1,..,n$,

and

μ_i is the associated dual (shadow) price, $i=1,..,n$.

This problem considers the total expected return associated with accepting walk-in requests for two guest-classes. We take first and second derivatives to show that $Z_{C1}(Y_i,Y_j)$ is concave and has an optimal solution such that $c_i\,[1-F(YS_i;Y_i)] = \mu_i$ and $c_j\,[1-F(YS_j;Y_j)] = \mu_j$. For $i < j$, the capacity allocated to class i can be downgraded to class j. So since we can always downgrade--but not upgrade--we want to keep $\mu_i \leq \mu_j$ and hence we get decision-rule (WALCON).

(WALCON)

For $i < j$, $j=2,..,n$, $[1-F(YS_i;YA_i)]/[1-F(YS_j;YA_j)] \leq c_j/c_i$

where

YS_i is the random variable for the number of walk-in's for the time remaining in the target period, $i=1,..,n$, and

YA_i is the limit on the number of class i walk-ins to accept, $i=1,..,n$. ∎

(WALCON) gives only limits on the relative sizes of walk-in request to accept. The absolute limits depend on net quantity of rooms available for walk-ins. This is deduced, with subjective judgements and given the service performance requirements, from the total capacity available, the number of booked reservations that remains on record, and the probability that they will show.

SALES MANAGEMENT AND RATES SETTING

We had assumed that the room rates are determined by competitive market forces. This is often true only for rack rates. To increase occupancy, hotels offer discounts to tour operators, convention organizers, and others. The hotel operators, therefore, have some discretion in setting the rates. The next rule provides some guidance on the relative value of rates for the guest-classes. It points out that the important contributors to rates differentials are the relative magnitudes of their reservation demand and survival characteristics.

We assume that the survival-rate distributions are independent of the demand distributions and consider the following problem.

(C2)

$$
\begin{aligned}
&Z_{C2}(N_i,N_j) \\
&= c_i \int_0^{N_i} \int_0^1 x\, N\, f(q_i;x) f(NS_i;N)\,dx\,dN + c_i\, N_i \int_{N_i}^{\infty} \int_0^1 x\, f(q_i;x) f(NS_i;N)\,dx\,dN \\
&+ c_j \int_0^{N_j} \int_0^1 x\, N\, f(q_j;x) f(NS_j;N)\,dx\,dN + c_j\, N_j \int_{N_j}^{\infty} \int_0^1 x\, f(q_j;x) f(NS_j;N)\,dx\,dN \\
&+ \pi_i(L_i - E[q_i]N_i) + \pi_j(L_j - E[q_j]N_j)
\end{aligned}
$$

where L_i is a given allocation of capacity to guest-class i, i=1,..,n

and π_i is the dual (shadow) price associated with the allocation, i=1,..,n.

This problem gives the total expected return associated to allocating the available capacities to two guest-classes. So we have a problem similar to the one for walk-in control. By taking first and second derivatives, it is easy to show that $Z_{C2}(N_i,N_j)$ is concave and has an optimal solution where $c_i\,[1-F(NS_i;N_i)] = \pi_i$ and $c_j\,[1-F(NS_j;N_j)] = \pi_j$.

Theorem 7: In the optimal solution for (C2), $\pi_i \geq \pi_j$ for $i < j$, $i=1,..,n-1$.

Proof of theorem 7: Suppose the theorem is false and $\pi_i < \pi_j$ for $i < j$. Then, we downgrade rooms from those allocated to class i to class j and gain an additional return of $(\pi_j - \pi_i)$ per unit downgraded. ∎

By the result presented in theorem 7, we give below the decision-rule for setting rates or granting discounts.

(RATESET)

For $i = 2,..,n$, $c_{i+1} \leq Q_i\, c_i$

where $Q_i = \dfrac{[1-F(NS_i;\, N_i]}{[1-F(NS_{i+1};\, N_{i+1})]}$, $i=1,..,n-1$. ∎

Here, we assumed that the relative values of the reservation targets, N_i, $i=1,..,n$ are given. (RATESET) suggests how market segmentation should be exploited: market should be segmented according to the strength of its demand relative to the availability of rooms. It also gives limits that will guide pricing negotiations with tour and convention groups. From above, for $i < j$, c_i is not always greater or equal to c_j. However, by our labelling convention, $c_i \leq c_j$ for $i < j$ when classes i and j are for the same room-type. But across room-types, guest classes in a room-type can have rates lower than the rack rate of a less luxurious room-type.

5. Comments and Extensions

The creation of the guest-class concept helps hotels earn more

revenue by exploiting market segmentation. It does so by controlling spills and diversions. Glover et al. [1982] gave the definition of spill and diversion for the airline context: "Spill is the movement of passengers to other flights, either the same or competing carriers. Diversion occurs when a passenger who would have stayed with the same carrier at the original higher fare takes advantage of a discount fare which was offered to stimulate increased occupancy, thus generating less revenue for the carrier." Spill, in our problem, refers to walk-in or reservation requests that the hotel has to turn away. We reduce spills from high-revenue guest-classes by controlling the number of low revenue requests to accept. Diversions are managed through better understanding of the characteristics of the market segments and applying to guests from these segments the appropriate rates.

The Parker house hotel in Boston actually created "service product" packages for different groups of customers that corresponds to what we have called guest-classes. The hotel's sales department pursue and develop the demand from these groups through direct contact. The capacity for tour group reservations are allocated after the capacity targeted to the higher paying groups have been accounted for, consistent with the outcome suggested by our analysis.

Airlines have been using authorization levels for reservations booking. The methods they used to obtain the authorization level are different from ours and they also do not account explicitly for downgrading effects. The airline reservations problem also deviates fundamentally from the hotel problem in that (except shuttle flights) it has fewer walk-ins. The alternatives available to the air-traveller are also restricted: the air traveller cannot just change to another flight when it has an oversale--there are very few flights that have the same destination and take off

within a short time of each other. Simple extensions can be made to apply our approach to the airline reservations problem.

On the other extreme, restaurants, like those famous seafood places in Boston, have so much demand that some do only walk-in business: they do not typically accept reservations. It is not difficult to provide a plausible explanation using the results of our analysis of the hotel problem: assuming other things being equal, holding reservations runs the additional risk of cancellations, late arrivals, and no-shows. Therefore, not only would there be situations when walk-in customers wait in frustration while tables lie idle, but the burden of management also increases.

New variations in the circumstances surrounding the problems like the penalty schemes to discourage no-shows: non-refundable sales, first day deposits, etc. are appearing. These present new challenges for extending our model which we leave for future research. Another area of future research is to explore the possible use of heuristics to solve the hotel problem. (DP2) has an interesting structure that suggests how one might work: a "knapsack" filling approach where we increase the values of decision variables that have the higher marginal returns first until the constraints are binding.

Finally, we will mention briefly how hotels measure their performance relative to each other. A common measure of operational efficiency for hotels is percent occupancy. One way of achieving high occupancy is to give large discounts and overbook excessively. Operating this way, the hotel fills up easily but reaps low revenue and, in violation of good practice, leaves many prospective reserved guests without rooms. Therefore, the level of occupancy does not fully reflect how well the hotel is managed.

Merliss and Lovelock [1980] highlighted an alternative performance measure (being used by the Parker House) called the room sales efficiency (RSE). RSE is the total room sales revenue over a period divided by the potential revenue that might be obtained if, during the same period, all available rooms were sold at rack rates. Maximizing expected return also maximizes expect RSE. This is an excellent measure for comparing hotels of different sizes and measuring how well they serve their market segments.

6. Summary and Conclusions

Previous studies consider the capacity allocation and the yield management problems independently. In this chapter, we show how they can be coordinated. We also showed how the profitability of a hotel can be optimized by careful utilization of its accommodation resources--not merely by increasing occupancy. The model we provide allows us to solve hotel reservations and sales planning problems that have multiple-day stays, multiple room-types, multiple guest-classes, and service constraints. We show that the problem can be separated into single-period problems. Using inner-linearization approximations, we can obtain near-optimal solution for the reservation targets. We also provide rules to assist in accepting walk-ins and in setting room rates. The rules can be applied to aid sales management and control discount offers. The model demonstrates, through the use of guest-classes, how the market segmented effectively can increase profits.

REFERENCES

GLOVER, F., R. GLOVER, J. LORENZO, and C. McMILLIAN 1982. "The Passenger-Mix problem in the Scheduled Airlines," Interfaces 12(3):507-520.

JENNINGS, J.B. 1981. "Booking Level Management," Proc. AGIFORS Symposium 1981.

KAO, E.P.C. and G.G. TUNG 1981. "Bed allocation in a Public Health Care Delivery System," Mgmt. Sci. 27:507-520.

KINBERG, Y., A.G. RAO, and E.F. SUDIT 1980. "Optimal Resource Allocation between Spot and Package demands," Mgmt. Sci. 26:890-900.

LADANY, S.P. 1976. "Dynamic Operating Rules for Motel Reservations," Decision Science 7:829-840.

LIBERMAN, V. and U. YECHIALI 1978. "On the Hotel Overbooking Problem - An Inventory System with Stochastic Cancellations," Mgmt. Sci. 24:1117-1126.

MERLISS, P.P. and LOVELOCK, C.H. 1980. "The Parker House: Sales and Reservations Planning," Harvard Business School case 9-580-152.

PFEIFER, P.E. 1989. "The Airline Discount Fare Allocation Problem," Decision Sci. 20(1):149-157.

RISING, E.J., R. BARON, and B. AVERILL 1973. "A Systems Analysis of a University-Health-Service Outpatient Clinic," Oper. Res. 21:1030-1047.

ROTHSTEIN, M. 1968. Stochastic Models for Airline Booking Policies, unpublished Ph.D. thesis, Graduate School of Engineering and Science, NYU, New York.

ROTHSTEIN, M. 1971a. "An Airline Overbooking Model," Trans. Sci. 5:180-192.

ROTHSTEIN, M. 1971b. "Airline Overbooking: The State of the Art," J. Trans. Econ. Policy 5:96-99.

ROTHSTEIN, M. 1974. "Hotel overbooking as a Markovian Sequential Decision Process," Dec. Sc. 5:389-404.

ROTHSTEIN, M. 1985. "OR and the Airline Overbooking Problem," Oper.Res. 33: 237-248.

SHLIFER, E. and Y. VARDI 1975. "An Airline Overbooking Policy," Transp. Sci. 9:101-114.

TAINTER, M. 1964. "Some Stochastic Inventory Models for Rental Situations," Mgmt. Sci. 11:316-326.

TAYLOR, C.J. 1962. "The Determinants of Passenger Booking Levels," Proc. the Second AGIFORS Symposium 1962:93-1161.

THOMPSON, H.R. 1961. "Statistical Problems in Airline Reservation Control," Opnl. Res. Quart. 12:167-185.

WHISLER, W.D. 1967. "A Stochastic Inventory Model for Rented Equipment," Mgmt. Sci. 13:640-647.

WILLIAM, F.E. 1977. "Decision Theory and the Inn Keeper: An Approach for setting Hotel Reservation Policy," Interfaces 7:18-30.

APPENDIX

NOTATIONS

n,m,T: Number of guest-classes, number of room-types, and length of planning horizon respectively.

$a(s)$: Smallest guest-class label for room-type s, $s=1,..,m$.

$A(s)$: Set of all guest classes for type s rooms. $A(s) = \{a(s),..,a(s+1)-1\}$.

$AU(s)$: Set of all guest-classes that can be offered type s rooms. $AU(s) = \{1,..,a(s+1)-1\}$, $s=1,..,m$.

c_{it}: Rate, per room per period, charged for guest-class i, $1=1,..,n$ in period t, $t=1,..,T$.

MU_{st}: Number of type s rooms available in period t.

M_{it}: Number of rooms initially allocated to guest-class i for period t. ($M_{it} = MU_{st}$ for $i=a(s)$ and $M_{it} = 0$ otherwise, $s=1,..,m$, $t=1,..,T$.)

W_{it}: Number of rooms from those allocated to class i to downgrade to class $i+1$ during period t and $W_{0t} = 0$, $t=1,..,T$.

q_{it}: Class i reservation show-rate (or survival-rate) for period t.

N_{it}: Number of reservations for class i guest in period t.

NS_{it}: Random variable for the demand of guest-class i reservations in period t.

$NS_{it}(y)$: $NS_{it}(y) = NS_{it}$ for $NS_{it} \leq y$ and $NS_{it}(y) = y$ otherwise, $i=1,..,n$ and $t=1,..,T$.

YS_{it}: Random variable for the number of class i prospective guests that will walk into the hotel during period t without reservations. $Y_{Sit} \in [0,\infty)$, $i=1,..,n$ and $t=1,..,T$, have finite mean and variance.

$YS_i(y)$: $YS_i(y) = YS_i$ for $YS_i \leq y$ and $YS_i(y) = y$ otherwise, $i=1,..,n$.

q,NS,YS: The vectors of random variables q_{it}, NS_{it}, YS_{it} respectively.

f(x;y): Probability density function of any random variable x evaluated at y.

F(x;y): Cumulative density function of any random variable x evaluated at y.

Prob(.): Probability of the event argument.

$E_x[.]$: Expectation over the random vector x.

α: Service performance target for booked reservations; probability target for meeting reservation demand. (Typically, $\alpha \in [0,1]$ is close to 1.)

$\phi(.)$: $\phi(a_1,..,a_n) = F^{-1}(\Sigma^n_{i=1}\, a_i\, q_i;\, \alpha)$ where $a_i \geq 0$, i=1,..,n.

J_{it}: $J_{it} = M_{it} + W_{i-1,t} - W_{it} - q_{it}\, N_{it}$, i=1,..,n and t=1,..,T.

J_{it}^+: $J_{it}^+ = \text{Max}(0,\, J_{it})$, i=1,..,n and t=1,..,T.

$J_{it}(y)$: $J_{it}(y) = M_{it} + W_{i-1,t} - W_{it} - q_{it}\, NS_{it}(y)$, i=1,..,n and t=1,..,T.

$J_{it}(y)^+$: $J_{it}(y)^+ = \text{Max}(0,\, J_{it}(y))$, i=1,..,n and t=1,..,T.

CHAPTER 5

DISTRIBUTION-FREE, UNIFORMLY-TIGHTER LINEAR APPROXIMATIONS FOR CHANCE-CONSTRAINED PROGRAMMING

1. Introduction

In this chapter, we present a class of linear approximations to chance-constrained problems where the random variables have arbitrary distributions. We focus on chance-constrained linear programs (LP) with stochastic technology coefficients--coefficients on the left-hand-side of a constraint. However, the method can be applied to mathematical programs that have chance-constrained linear or nonlinear inequalities with stochastic or deterministic resource parameters, the right-hand-side parameter of a constraint. These problems appear in service constrained applications that have uncertainties in the yield or demand. The broad classification of these applications are problems (a) with carry-over resources (inter-period constraints), (b) with portfolio selection (intra-period constraints), and (c) with both carry-over resources and portfolio selection.

Applications of the first type can be found in the areas of production planning and inventory control; facility location planning; project planning (PERT); financial investment planning; cash management; cost-volume-profit analysis; and environmental, public services, and utilities (hospital staffing, reservoir capacity, rail-road system, solid waste system) planning. Portfolio-selection applications include problems in investment portfolio management, activity analysis and technology planning, capital budgeting, animal and human dietary planning, and material composition selection. The lists are not intended to be

comprehensive but illustrates the variety of applications. For more applications and detail of specific problems, see [Hogan, Morris, and Thompson 1981] and the papers listed in our reference.

In general, chance-constrained programs with random technical coefficients are hard to solve. To make the problem tractable, it is typical to assume that the random coefficients are normally distributed. We propose, in this chapter, a new alternative approach. The method relaxes the assumptions needed for the probability distributions of the random coefficients. Our main goal is to derive an approach that is intuitive and allows easy extraction of a problem's structural properties for building simple heuristics. Examples of applying the method appear in chapters 2 through 4 of this thesis.

The chapter is organized as follows. Section 2 review the work of other researchers. Section 3 describes the general principle of our approach and illustrates with some examples. In section 4, we recapitulate the equations used in the major alternative approximations and test our method against them. We report the results in section 5 with comments on extensions and future research. In the last section, we conclude with a short summary.

2. Literature Review and Important Results

(For a quick tutorial on stochastic programming problems and chance-constrained programs, we refer the reader to the introductory comments in [Hillier 1967] and the references mentioned there. Greater details on the subject can be found in stochastic programming texts like [Sengupta 1972], [Vadja 1972], [Kall 1976], and [Dempster 1980].)

Under the condition that the joint distribution function of the random variables is continuous, Symonds [1967] proved that the feasible

region of a chance-constrained program and its deterministic equivalent coincide. He also provided results for linear chance-constrained problems where only the resource vector is random. Problems with random resource is quite well covered in the literature. In fact, most of the research on chance-constrained problems have centered on the cases where the resource only is random. The reader is directed to see [Charnes, Cooper, and Symonds 1958], [Charnes and Cooper 1963], and others listed in our reference for more details. In a LP with stochastic technology coefficients, a chance constraint comprises a linear combination of random variables. The weights used for combining the random variables are the decision variables of the program. Hence in general, before solving the program, the probability of violating the resource capacity is difficult to evaluate.

To simplify discussion, without loss of generality, we assume that the resource parameter is deterministic. Problems with stochastic resource parameters in the chance constraints can be transformed by multiplying the parameters with dummy decision variables, thus converting them into technology coefficients. Adding new constraints to the problem, we set the value of these dummy variables to one. Now consider the chance-constrained linear inequality

$$\text{Prob}(\Sigma^{n}_{i=1}\ a_i\ x_i \leq b) \geq \alpha. \qquad (1)$$

a_i, i=1,..,n, are the random technology coefficient variables with continuous joint distribution; x_i, i=1,..,n, are the decision variables; b is the deterministic resource parameter; and $\alpha \in [0,1]$, the service performance target, is the probability that $\Sigma^{n}_{i=1}\ a_i\ x_i \leq b$ is satisfied. The random variables a_i, i=1,..,n, are assumed to have finite mean $E[a_i]$ and finite variance $V[a_i] = \sigma_i^2$.

By Symonds' theorem, constraint (1) has a deterministic equivalent

$$g(x) \leq b, \tag{2}$$

with vector $x \equiv (x_1,..,x_n)$. In general, $g(x)$ is a nonlinear function and can be written as

$$g(x) = E[\Sigma^n_{i=1}\ a_i\ x_i] + z_\alpha\ (V[\Sigma^n_{i=1}\ a_i\ x_i])^{1/2}, \tag{3}$$

where $z_\alpha = z_\alpha(x)$, the safety factor, is a function of the service target α as well as the decision vector x. When a_i, $i=1,..,n$, are independent and normally distributed, (3) becomes

$$g(x) = \Sigma^n_{i=1} E[a_i]\ x_i + Z_\alpha\ [\Sigma^n_{i=1}\ \sigma_i^2\ x_i^2]^{1/2}\ . \tag{4}$$

Here Z_α is the "one-tail" normal variate for α and depends on α only. This value can be obtained easily from the tables in most basic statistics texts. (4) is true because a linear combination of normal random variables is also normally distributed.

The representation in (4) is also applicable to the class of stable probability distributions (see [Allen, Braswell, and Rao 1974]). Stable distributions are distributions, completely specified by their means and standard deviations, such that a linear combination of random variables with a common stable distributional form have the same distributional form. It is not necessary for the random variables to be identically distributed. Their means and variances may be different but they must be from a common distributional form. The class of stable distributions include the normal, Poisson, Chi-square, and binomial distributions. Though not as readily available as the normal variate, the safety factor for the other stable distributions can be computed or found in published tables. Again, the safety factor for stable distributions is independent of the decision variables: the safety factor is a function of α only.

CONDITION C1: Random variables a_i, i=1,..,n, shares a common stable distributional form. ▪

When condition C1 is true and the distributions are dependent, z_α is still a function of α only. However, in the general, since z_α may be dependent on x, g(x) is difficult to evaluate and usually cannot be expressed in closed form.

THEOREM 1 [Kataoka 1963]: A chance-constrained linear inequality under C1 and with $z_\alpha \geq 0$ is convex. ▪

Theorem 1 is the main published result on the convexity of a chance-constrained linear inequality. Little is known to date about the convexity of (1) when the random variables are of other distributions. Hillier [1967] mentioned that for arbitrary distributions, under fairly weak conditions, the central limit theorem may permit the normal approximation. Charnes, Cooper, and Thompson [1963] suggested that a mixture of normal distributions can be used to approximate distributions of fairly arbitrary shapes. Under this approximation, the term $\Sigma^n_{i=1}\ a_i\ x_i$ in (1) is again a normal random variable. Z_α, the safety factor for the normal distribution, is non-negative when $\alpha \geq 0.5$. Therefore, using theorem 1, we can argue that for $\alpha \geq 0.5$, (1) is usually convex. Our simulation experiments, on common distributions, indicate that for α close to 1, (1) is convex. We did not find any general theoretical result on the neccessary conditions for convexity of (1). Some work in this area have been done by Prekopa [1971,1974]. This remains an interesting research question for further study.

ASSUMPTION A1: The feasible region of (1) is convex. ▪

We will assume A1 to be true for the rest of this chapter. This condition has been assumed to be true in almost all the work we came across. Hence

the assumption we make is no more restricted than those that have been made (see for example, [Hillier 1967] and [Seppälä 1971]).

The usual approach taken by previous studies, after assuming C1 and convexity, is usual to solve the chance-constrained problem with nonlinear programming methods. The nonlinear programming methods usually linearize the problem and search along subgradients. An example of this is Kelly's [1960] cutting plane method. This method solves chance-constrained programs, in multiple passes, as linear programs. A linear program is first solved without the chance constraints. At each subsequent iteration, a hyperplane tangent to each chance constraint is defined using the preceding iteration's solution and its partial derivatives. These are introduced into the program as additional linear constraints. Prekopa [1988] summarizes the numerical approaches available for solving chance-constrained problems. These nonlinear programming methods tend to be complicated: they require partial derivatives, multiple-pass techniques, non-standard computer codes. Moreover, they are usually restricted to cases under condition C1.

Allen, Braswell, and Rao [1974] developed methods for approximating a chance-constrained set using information derived from sample data only. This approach is based on Wilks's [1963] work on the use of statistically equivalent blocks to construct 100α% tolerance regions. The level of confidence of satisfying the chance constraints can be determined from the size of the sample. In addition to the type-one error information implicit in the chance-constraints, this approach gives the decision-maker type-two error information. The paper compared the percentages of empirical constraint satisfaction of the actual feasibility region by (a) assuming stable distribution, (b) using a safety factor derived from Chebyshev's inequality, (c) using a simple linear approximation, and (d) using a

hyperspherical approximation. Charnes, Kirby, and Raike [1970] studied this further and developed the "acceptance region theory".

Allen, Braswell, and Rao's approximations are not uniformly tighter; the solutions to their approximations are not necessarily feasible to the original problem. But the uniformly-tighter characteristic is important in many practical applications of chance-constrained programs. So Hillier [1967] and later Seppälä [1971, 1972] devised approaches that are uniformly tighter. Both methods require condition C1. In this section, we give a brief sketch of these two methods. We will provide their technical detail in section 4.

Hillier restricted his study to cases where the decision variables are (a) 0-1, (b) 0 or 1, or (c) bounded. He approximated the variance term in (3) with a separable nonlinear function. The separability property of the approximation makes it easier to apply nonlinear programming techniques. The nonlinear approximation can also be further approximated with a piecewise-linear function using a standard approach in separable convex programs ([Bradley, Hax, and Magnanti 1977]). The piecewise-linear approximation is uniformly tighter than the nonlinear approximation; in turn, the nonlinear approximation is uniformly tighter than the chance constraint.

Despite the set-back of not getting the exact optimal solution, Hillier expounded on the value of linear approximations. The advantages he listed are (a) the relatively high efficiency of solving linear programs, (b) the ability to do sensitivity analysis, and (c) the availability of linear duality theory for analyzing the solutions for managerial implications. The service levels in the chance constraints may be initially selected by managerial policy. After solving the problem, the service levels should be re-evaluated against their corresponding optimal dual

variable values. These give a measure of the costs of maintaining these service levels and hence provide guidance for revising them. Hence, duality results are important.

Seppälä [1971] relaxed the restrictions, on the decision variables, required by Hillier. Focussing, as in Hillier's approach, on the variance term of the deterministic equivalent (3), Seppälä introduced new variables to break this nonlinear variance term into simpler nonlinear functions. Each nonlinear function is a function of two variables only. He then approximated each of these by piecewise-linear segments. The resulting linear approximation is uniformly tighter than the chance constraint. In this, as well as Hillier's approach, the decision-maker can choose the number of assisting linear constraints. As the number of linear constraints increases, the error of Seppälä's solution from the optimum approaches zero. For Hiller's method, the error will decrease with more linear constraints but may not go to zero.

In a recent paper, Olson and Swenseth [1987] suggested the simple approximation of replacing each random coefficient a_i, i=1,..,n, in (1) by the sum of its expected value and its standard deviation multiplied by a safety factor, shown below:

$$g(x) = \Sigma^{n}_{i=1}(E[a_i] + z_\alpha \sigma_i) x_i \ . \qquad (5)$$

Assuming C1 and with the means and the variances given, this approximation is the same as Allen, Braswell, and Rao's [1974] linear approximation. As an illustration, Olson and Swenseth solve Van de Panne and Popp's [1963] cattle-feed mix problem--assuming independent normal distributions--and showed that the errors are small.

This linear approximation consists of only one inner-linearization hyperplane. The hyperplane defines a half space that guarantees feasibility to (1). Since the feasible region defined by a chance constraint is

nonlinear, in the worst case the gap from optimality can be very large. This is particularly so when the magnitudes of the variances relative to the means of the random variables are huge. Fortunately for them, the coefficients of variation (COV) in the cattle-feed mix problem are extremely small--less than 0.01--and hence the excellent results.

In cases where condition C1 is not true, both Hillier and Seppälä suggest that the Chebyshev's inequality be used to obtain the safety factor. However, the Chebyshev's inequality is known to give safety factors with magnitudes much larger they need to be and hence the approximations using them will excessively constrict the feasible region (e.g. see Allen, Braswell, and Rao [1974]). They are particularly bad when the random variables have finite supports (for example, $a_i \geq 0$ or $a_i \in [0,1]$). In the literature we have encountered, most test cases have small COV's. However, in some applications the COV's can be large: Albin and Friedman [1989] reported that the distributions of defects in integrated circuit fabrication have COV's larger than 1.

3. Description of the method

Unlike Hillier and Seppälä who linearize the variance term in (3), we linearize g(x) in (2). This way, we do not have to deal with z_α explicitly. Our method constructs hyperplanes, each formed by connecting selected points on the boundary of (2). The hyperplanes collectively approximate the feasible region of (2). The main principle in our approach is to make some "guesses" about the relative magnitudes of the decision variables. For each guess, g(x) becomes a function with one random coefficient and one decision variable. Specifically, we "guess" that $x_i = s_i w$, $i = 1,..,n$, for a selected deterministic vector $s \equiv (s_1,..,s_n)$ and decision variable w. In this case, (1) becomes

$$\text{Prob}(w \Sigma^{n}_{i=1} a_i s_i \leq b) \geq \alpha. \quad (6)$$

Constraint (6) can be re-written as linear inequality

$$\phi(s)\, w \leq b \quad (7)$$

where fractile $\phi(s) = F^{-1}(\Sigma^{n}_{i=1} a_i s_i; \alpha)$ is a deterministic coefficient, $F(u;v)$ is the cumulative density function of random variable u evaluated at v, and $F^{-1}(u;.)$ is the inverse function of $F(u;v)$.

Vector s corresponds to a ray from the origin. With each vector s, the evaluation of (7) gives us a point on the upper boundary surface of (2). We repeat the process for a set of selected rays to get a set of points on the boundary of (2). We then connect adjacent points to form hyperplanes. The hyperplanes are introduced into the problem as linear inequalities (linear constraints). These linear inequalities replace the chance constraint (1) in the problem.

Geometrically, the linear inequalities form a polyhedron that has extreme points touching the upper boundary of the feasible region. Consequently, if the feasible region formed by (2) is convex, it "contains" the polyhedron. Therefore when (2) is convex, the set of linear inequalities is uniformly tighter than (2). The extreme points are the same points at which the selected rays from the origin intersect the upper boundary of the feasible region. Figure 1 illustrates the polyhedron formed by the extreme points in a convex chance constraint in R^2.

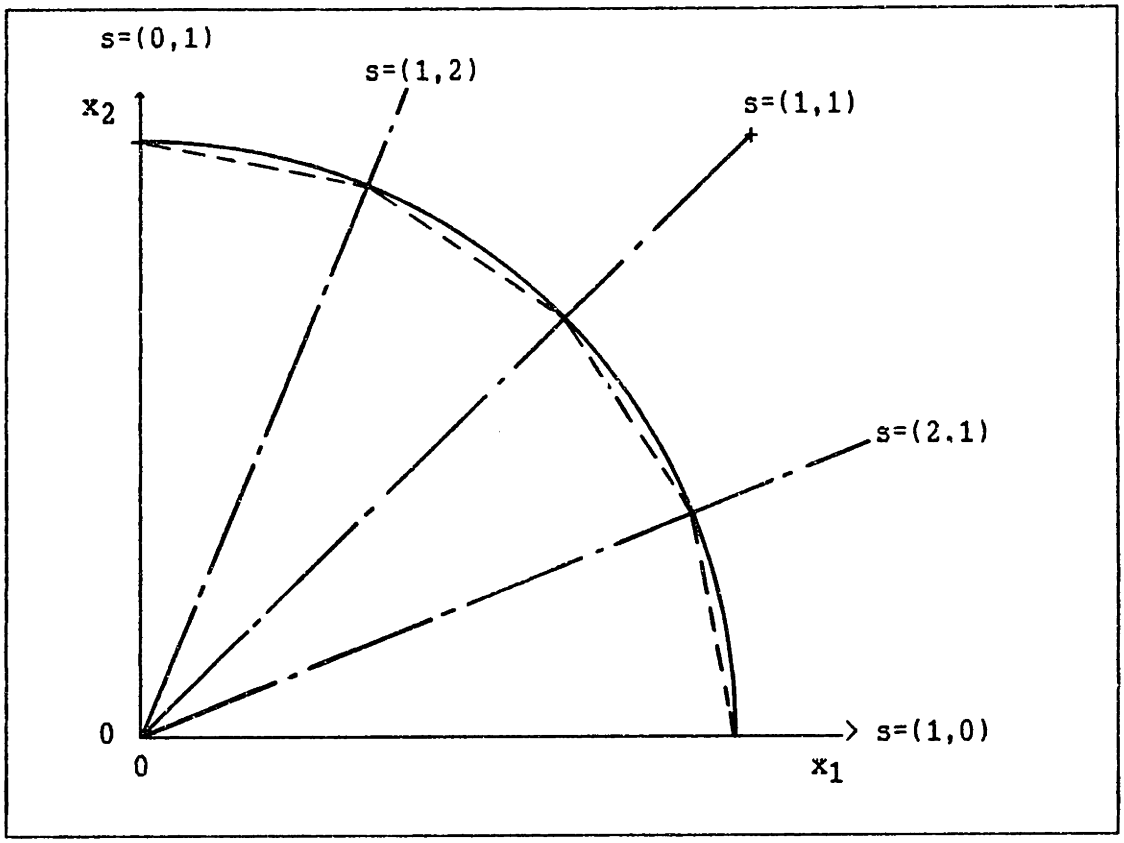


Figure 1. Linear Approximation of a Chance Constraint.

With an infinite number of rays, the linear inequalities reproduce the chance constraint (1). In practice, only a few well-chosen rays are needed to give solutions with small relative errors from the optimal value.

EXAMPLES

All our linear approximations have the form

$$\Sigma^{n}_{i=1}\ \Omega_{ik}\ x_i \leq b, \quad k=1,..,K, \tag{8}$$

where Ω_{ik}, i=1,..,n, are deterministic coefficients. For each k, $\Sigma^{n}_{i=1}\ \Omega_{ik}\ x_i = b$ defines a hyperplane and Ω_{ik}, i=1,..,n, are obtained by solving a system of equations; each equation corresponds to an extreme point (of the polyhedron) that rests on the hyperplane. This effort is done once only for each approximation; they are not solved each time the approximation is used.

In this study we provide, as examples of the general approach, three approximations: (RAY1). (RAY2), (RAY3). The rays used in these

approximations are the decision variable axes and the centriod rays, rays in the center of the cone formed by subsets of the axes. Allen, Braswell, and Rao [1974] and Olson and Swenseth's [1987] methods are special cases of (RAY1); the rays used in (RAY1) are the decision variable axes only.

We define the unit vector $e_i = (s_1,..,s_n)$ where $s_j = 1$ for $j = i$ and $s_j = 0$ otherwise. Below, we provide the approximations.

(RAY1)

$g_{RAY1}(x) = \Sigma^n_{i=1}\ \Omega_{ik}\ x_i,\quad k=1,..,K=1,$

where $\Omega_{ik} = \phi(e_i),\ k=1,..,K=1.$

(RAY2)

$g_{RAY2}(x) = \Sigma^n_{i=1}\ \Omega_{ik}\ x_i,\quad k=1,..,K=n,$

where $\Omega_{ik} = \phi(\Sigma^n_{j=1}\ e_j) - \Sigma^n_{j=1}\ \phi(e_j) + \phi(e_i)$ for $i = k$

and $\Omega_{ik} = \phi(e_i)$ otherwise.

For n = 3, for example, matrix $\Omega \equiv [\Omega_{ik}]$

$$= \begin{bmatrix} \phi(1,1,1)-\phi(0,1,0)-\phi(0,0,1) & \phi(0,1,0) & \phi(0,0,1) \\ \phi(1,0,0) & \phi(1,1,1)-\phi(1,0,0)-\phi(0,0,1) & \phi(0,0,1) \\ \phi(1,0,0) & \phi(0,1,0) & \phi(1,1,1)-\phi(1,0,0)-\phi(0,1,0) \end{bmatrix}$$

For (RAY3), we construct unique sets $\{\pi(i,k),\ i=1,..,n\},\ k=1,..,K.$, By itself, each $\pi(i,k)$ is also a set such that $\pi(i-1,k) \subset \pi(i,k)$, $i=1,..,n$, with $\pi(n,k) = \{1,..,n\}$ and $\pi(0,k) = \{\}$ for all k. Note that there can be n! unique sets of $\{\pi(i,k),\ i=1,..,n\}$ and hence K = n!.

(RAY3)

$g_{RAY3}(x) = \Sigma^n_{i=1}\ \Omega_{ik}\ x_i,\quad k=1,..,K=n!,$

where $\Omega_{ik} = \phi(\Sigma_{r\in\pi(j,k)}\ s_r) - \phi(\Sigma_{r\in\pi(j-1,k)}\ s_r)$

with j such that $i = \pi(j,k)\backslash\pi(j-1,k)$. (\ is the set subtraction operation.)

Now for n = 3,

$$\Omega = \begin{bmatrix} \phi(1,1,1)-\phi(0,1,1) & \phi(0,1,1)-\phi(0,0,1) & \phi(0,0,1) \\ \phi(1,1,1)-\phi(0,1,1) & \phi(0,1,0) & \phi(0,1,1)-\phi(0,1,0) \\ \phi(1,0,1)-\phi(0,0,1) & \phi(1,1,1)-\phi(1,0,1) & \phi(0,0,1) \\ \phi(1,0,0) & \phi(1,1,1)-\phi(1,0,1) & \phi(1,0,1)-\phi(1,0,0) \\ \phi(1,1,0)-\phi(0,1,0) & \phi(0,1,0) & \phi(1,1,1)-\phi(1,1,0) \\ \phi(1,0,0) & \phi(1,1,0)-\phi(1,0,0) & \phi(1,1,1)-\phi(1,1,0) \end{bmatrix}$$

(RAY1), (RAY2), and (RAY3) are by no means the only approximations possible using our methodology. There can be numerous variations by using different rays and constructing different hyperplanes from them. Moreover, an iterative multi-pass approach may be devised to generate the rays based on previous iterations solution. We leave this challenge for future research.

FRACTILES

When the joint probability distribution of the random variables are known, the fractiles $\phi(s)$ sometimes may be obtained as closed-form expressions. Alternatively, at a level of confidence that corresponds to their sample sizes, the fractiles can be extracted from Monte-Carlo simulation (Levy [1967]) or sample data. If the form of the distribution is known but not the exact distribution, the fractiles may be estimated using statistical approximation methods (Bache [1979], Cornish and Fisher [1937], Fisher and Cornish [1960]). These methods can make use of the sampling information beyond the first two moments. When the sample data only is available, distribution-free or non-parametric methods (Allen, Braswell, and Rao [1974], Wilks [1963]) may be used.

4. Comparative Experiments

We compare our method against those by Hillier, Seppälä, and Olson and Swenseth. Before proceeding, we sketch how these methods approximate g(x) for the test conditions we are using. Following that, we describe the test conditions and then state the comparison criterion.

$g(x) \equiv F^{-1}(\Sigma^n_{i=1}\ a_i\ x_i;\ \alpha)$

Normal distribution:

$g(x) = \Sigma^n_{i=1}\ E[a_i]\ x_i + Z_\alpha\ [\Sigma^n_{i=1}\ \sigma_i^2\ x_i^2]^{1/2}.$

Uniform distribution and $x_i \in \{0,1\}$, i=1,..,n:

$g(x) = m - (m!(1-\alpha))^{1/m}$ where $m = \Sigma^n_{i=1}\ x_i$.

For $x_i \in \{0,1\}$, g(x) for the uniform distribution case can be derived using geometry.

(HILL)--Hillier

Normal distribution and $x_i \in [0,1]$, i=1,..,n:

$g_{HILL}(x) = \Sigma^n_{i=1}\ E[a_i]\ x_i + Z_\alpha\ [\Sigma^n_{i=1}\ (\theta^2 - \sigma_i^2 + \sigma_i^2\ x_i^2)^{1/2} - (n-1)\theta]$

Uniform distribution and $x_i \in \{0,1\}$, i=1,..,n:

$g_{HILL}(x) = \Sigma^n_{i=1} E[a_i]\ x_i + (\alpha/(1-\alpha))^{1/2}\ [\Sigma^n_{i=1}\ (\theta - (\theta^2 - \sigma_i^2)^{1/2})\ x_i + \Sigma^n_{i=1}\ (\theta^2 - \sigma_i^2)^{1/2} - (n-1)\theta]$

where $\theta = (\Sigma^n_{i=1}\ \sigma_i^2)^{1/2}$.

$(\alpha/(1-\alpha))^{1/2}$ is the one-sided Chebyshev's inequality safety factor. Approximation $g_{HILL}(x)$ given above is only the nonlinear approximation of Hillier's method. For the comparison tests, we did not linearize it. The reader will realize that, for our comparison criterion, the result using $g_{HILL}(x)$ will be an upper bound on the linearized version.

(SPLK)--Seppälä

Normal distribution and $x_i \geq 0$, i=1,..,n:

$$g_{SPLK}(x) = \Sigma^n_{i=1} E[a_i]\, x_i + Z_\alpha\, y_n$$

where $y_i = \text{Max}\{r_{ik}\, y_{i-1} + s_{ik}\, x_i,\ k=1,..,K\}$ with $y_0 = 0$,

$$r_{ik} = \frac{t_{ik}\,(1 + \sigma_i^2\, t_{i,k-1}^2)^{1/2} - t_{i,k-1}\,(1 + \sigma_i^2\, t_{ik}^2)^{1/2}}{t_i - t_{i,k-1}},$$

$$s_{ik} = \frac{(1 + \sigma_i^2\, t_{i,k}^2)^{1/2} - (1 + \sigma_i^2\, t_{i,k-1}^2)^{1/2}}{t_i - t_{i,k-1}},$$

and $t_{ik} = \text{TAN}[(k/K)(\pi/2)]/i^{1/2}$, i=1,..,n, k=1,..,K.

Uniform distribution and $x_i \geq 0$, i=1,..,n:

Same as above except replace Z_α by $(\alpha/(1-\alpha))^{1/2}$.

We consider the instances where K = 3 and 6 which we label (SPL3) and (SPL6) respectively. Seppälä's approach replaces each chance constraint with O(nK) linear inequalities.

(OLSW)--Olson and Swenseth

Normal distribution and $x_i \geq 0$, i=1,..,n:

$$g_{OLSW}(x) = \Sigma^n_{i=1} (E[a_i] + Z_\alpha\, \sigma_i)\, x_i.$$

Uniform distribution and $x_i \geq 0$, i=1,..,n:

$$g_{OLSW}(x) = \Sigma^n_{i=1} (E[a_i] + (\alpha/(1-\alpha))^{1/2}\, \sigma_i)\, x_i.$$

(OLSW) is the same as (RAY1) under condition C1 and if z_α is available.

(RAY1, RAY2, and RAY3)

The approach we propose in this study gives not one approximation but a class of approximations. But for the purpose of comparison, we use only (RAY1), (RAY2), and (RAY3). Under the normal distribution, $\phi(s) = \Sigma^n_{i=1} E[a_i]\, s_i + Z_\alpha\, [\Sigma^n_{i=1} \sigma_i^2\, s_i^2]^{1/2}$; and for the uniform distribution case with $s_i \in \{0,1\}$, $\phi(s) = m - (m!(1-\alpha))^{1/m}$ where $m = \Sigma^n_{i=1} s_i$. With the

expression for $\phi(s)$ given, approximations (RAY1), (RAY2), and (RAY3) are completely defined.

For simplicity of conducting the experiments, we test only cases where the random variables are independent identically distributed. All the methods tested can be used for dependent and non-identically distributed random variables. First, we evaluate cases where a_i, i=1,..,n, are normally distributed with $x_i \in [0,1]$, i=1,..,n. Here, we examine how well our methods compare against the other methods in the conditions specified for those methods. Then we compare the methods when a_i, i=1,..,n, are uniformly distributed with $x_i \in \{0,1\}$, i=1,..,n. In general, the conditions are specified to keep the tests simple or to replicate the conditions that were originally intended for the other methods.

The test input data are generated as follows:

(a) For the normal distribution cases, we let $E[a_i] = 0.5$, i=1,..,n, and pick variance $\sigma_i^2 = \sigma^2$, i=1,..,n, from a uniform distribution in [0,VA]. We examine cases where $VA = 10^{-\beta}$, $\beta = 0,..,4$. The maximum coefficient of variation (COV), in each case of β, is $\sqrt{VA}/0.5$. Overall, the COV ranges from 0 to 2. The values of decision variables x_i, i=1,..,n, are sampled from a uniform distribution in [0,1]. We tested cases where n = 2, 4, and 6.

(b) For the uniform distribution cases, we sample a_i, i=1,..,n, from a uniform distribution in [0,1]. Hence, $E[a_i] = 0.5$, i=1,..,n, and variance $\sigma_i^2 = \sigma^2 = 1/\sqrt{12}$ and $COV = (1/\sqrt{12})/0.5 = 0.58$. To obtain $x_i \in \{0,1\}$, we sample from a uniform distribution in [0,1]. We then let $x_i = 0$ when the sampled value is less than 0.5; and $x_i = 1$ otherwise. We evaluated the cases where n = 2, 4, 6, 8, 16, and 32.

The criteria of comparison is the relative error that the values of the approximation deviate from g(x): relative error of method h $\equiv (g_h(x)-$

$g(x))/g(x)$. Since all the methods to be compared are uniformly tighter than (1), $g(x) \leq g_h(x) \leq b$, $h \in$ {HILL, SPL3, SPL6, OLSW, RAY1, RAY2, RAY3}. Therefore, relative error is non-negative; it is a measure of how constricted approximation h is when (1) is binding and the sampled value of x is the optimum solution. We replicated each set of test conditions 15 times and compute their average and standard deviation.

5. Results and Comments

First, we notice that our three approximations have different computation requirements. (RAY1) replaces each chance constraint with one linear inequality; (RAY2) uses n linear inequalities and (RAY3) uses n! linear inequalities. (RAY1) and (RAY3) represent the extreme types of our class of approximations. (RAY3), especially, will be difficult to implement for n larger than 6; it took about 50 seconds to evaluate one chance-constrained approximation on a IBM compatible 80286 class machine when n = 6. (RAY1), (RAY2), and other methods took a negligible amount of time--typically less than 1 second per evaluation of g(x). Second, the result of (RAY3) dominates that of (RAY2) and the result of (RAY2) dominates that of (RAY1). This is because (RAY3) is uniformly tighter than (RAY2) and (RAY2) is uniformly tighter than (RAY1). We can therefore use the results of one approximation as a bound to another.

The comparisons outcome is tabulated in figures 2a through 3b. Figure 2a tabulates the average and standard deviation of the relative errors for the normal distribution cases. These results are graphed in figures 2b through 2d, for n = 2, 4, and 6 respectively. For maximum-COV less than 0.5, (RAY2) does as well as (HILL) and (SPL6), but it gets progressively worse as n becomes larger. The average relative errors in such cases are about 5% or less. Since (SPL6) uses 6 times more constraints than (RAY2)

and (HILL) is only the nonlinear approximation part of Hillier's method, the comparable results among the three methods suggest that our method performs very well.

[INSERT FIGURES 2a THROUGH 2d HERE]

For the uniform distribution cases, figure 3a tabulates the results while figure 3b presents it graphically. In these cases for n up to 6, (RAY3) is exact. We did not test (RAY3) beyond n = 6 since the computation effort is too much. (RAY1) and (RAY2) performed equally well and are about an order of magnitude better than (HILL) and (SPL6) when n is less than 6. They dominate the other methods up to n = 16. The methods, except for (OLSW), have about the same size of errors when n is larger than 16. (OLSW), on the whole, demonstrated to be an inferior method.

[INSERT FIGURES 3a AND 3b HERE]

The results of both the normal and uniform distribution cases show that the average relative errors become worse with larger n. As such, we do not recommend using any of the methods, ours included, for COV's larger than 0.5 when n gets beyond 12; the average relative errors then goes up to more than 50%. However, in some problems, the structural properties of the problem may permit large n. For example, in production/inventory problems, the service constraints for the planning horizon have a block triangular structure. The chance constraints with few variables are the service constraints for the more immediate periods; the periods further into the horizon have more variables. Thus when the planning is done on a rolling horizon, the large errors for the periods further away are not important as

long as the nearer periods are well approximated. This will be the case when a method like (RAY2) is applied. In other instances, capacity constraints may limit the number of non-zero decision variables. Again, (RAY2) should perform very well there.

EXTENSIONS AND FUTURE RESEARCH

We have assumed zero-order decision rules. The linear formulation of our method can be adapted easily to give solutions for linear decision rules. Therefore for problems with sufficient stationarity, we solve the problem once, replacing all decision variables by decision rules. The decision rules which are functions of the system state can then be applied with the immediate system state as input.

The chance constraints we have examined are for linear inequalities--that is, the term inside of Prob(.) in (1) is a linear function. Our approach does not restrict us to linear inequalities; we can have nonlinear inequalities or situations where $a_i = a_i(x)$, i=1,..,n, are functions for x. In production problems with stochastic yield, the last situation corresponds to the case where the yields are not independent of the lot size; a problem that has chance constraint like constraint (1) corresponds to the case where the yields are independent of the lot size.

For future research, we hope to consider the following: (a) examine how the relative errors can be parametrically bounded, (b) provide a multi-pass ε-optimal iterative approach, (c) consider problems with joint chance constraints, (d) provide theoretical results on the convexity of the feasible region.

6. Summary and Conclusions

We provided a class of linear approximations for problems with individual chance constraints that have random variables with arbitrary

distributions. In our method, we linearized the nonlinear deterministic equivalents of chance constraints when the deterministic equivalents may not have closed-form expressions. The linear approximation is uniformly tighter than the chance constraint when the feasible region defined by the chance constraint is convex. Therefore under convexity, the solutions generated by our approximations will satisfy or do better than the service target specified by the chance constraint.

Our method gives linear inequalities that retain the original decision variables. This make it easy to extract heuristics or apply higher order decision rules. The resulting programs are linear programs which permit the use of standard LP codes, perform sensitivity analysis, and have both primal and dual solutions. The coefficients in the approximating linear constraints can be extracted from assumed distributions or sample data. We do not restrict the distributional form of the random variables, disallow their dependencies on each other or require their partial derivatives.

The user has the option of trading-off the amount of computation effort against the accuracy of the solution by selecting the number of linear inequalities. In the simulation tests, applications of our approach with very few inequalities compare very well against the existing methods for both the normal and the uniform distributions.

REFERENCES

ALBIN, S.L. and D.J. FRIEDMAN 1989. "Impact of Clustered Defect Distributions in IC Fabrication," Mgmt. Sci. 35:1066-1078.

ALLEN, F.M., R. BRASWELL, and P. RAO 1974. "Distribution-free Approximations for Chance-Constraints," Op. Res. 22:610-621.

BACHE, N. 1979. "Approximate Percentage Points for the Distribution of a Product of Independent Positive Random Variables," Appl. Statist. 28:158-162.

BRADLEY, HAX, and MAGNANTI 1977. Applied Mathematical Programming, Addison-Wesley Publ., Reading, Mass.

CHARNES, A., W.W. COOPER, and G.H. SYMONDS 1958. "Cost Horizons and Certainty Equivalents: an Approach to Stochastic Programming of Heating Oil," Mgmt. Sci. 4:235-263.

CHARNES, A., and W.W. COOPER 1963. "Deterministic Equivalents for Optimizing and Satisficing under Chance Constraints," Op. Res. 11:18-39.

CHARNES, A., W.W. COOPER, and G.L. Thompson 1962. "Characterizations by Chance-constrained Programming," in Recent Advances in Mathematical Programming, Graves, R. and P. Wolfe (editors).

CHARNES, A., M. KIRBY, and W. RAIKE 1970. "An Acceptance Region Theory for Chance-constrained Programming," J. Math. Anal. Appl. 32:38-61.

CORNISH, E.A. and R.A. FISHER 1937. "Moments and Cumulants in the Specification of Distributions," Rev. Int. Statist. Inst. 5:307-321.

DEMPSTER, M.A.H. 1980. Stochastic Programming, (Dempster ed.), Academic Press, New York.

FISHER, R.A. and E.A. CORNISH 1960. "The Percentile Points of Distributions having Known Cumulants," Technometrics 2:209-225.

HOGAN, A.J., J.G. MORRIS, and H.E. THOMPSON 1981. "Decision Problems under Risk and Chance-Constrained Programming: Dilemmas in the Transition," Mgmt. Sci. 27:698-716.

HILLIER, F.S. 1967. "Chance-Constrained Programming with 0-1 or Bounded Continuous Decision Variables," Mgmt. Sci. 14:34-57.

KALL, P. 1976. Stochastic Linear Programming, Springer-Verlag, New York.

KATAOKA, S. 1963. "A Stochastic Programming Model," Econometrica 31:181-196.

KELLY, J.E. 1960. "The Cutting Plane Method for Solving Convex Programs," SIAM Jour. 8:703-712.

LEVY, L.L. and A.H. MOORE 1967. "A Monte-Carlo Technique for obtaining System Reliability Confidence Limits from Component Test Data," IEEE trans. Rel. R-16:69-72.

OLSON, D.L. and S.R. SWENSETH 1988. "A Linear Approximation for Chance-Constrained Programming," J. Opl. Res. Soc. 38:261-267.

PREKOPA, A. 1971. "Logarithmic Concave Measures with Applications to Stochastic Programming," Acta Sci. Math. (Szeged) 32:301-316.

PREKOPA, A. 1974. "Programming under Probabilistic Constraints with a Random Technology Matrix," Math. Operationsforsh. Statist. 5:109-116.

PREKOPA, A. 1988. "Numerical Solution of Probabilistic Constrained Programming Problems," in Numerical Techniques for Stochastic Optimization (Y. Ermoliev and R. J-B. Wets eds.), Springer-Verlag, New York.

SENGUPTA, J.K. 1972. Stochastic Programming--Methods and Applications, North-Holland.

SEPPÄLÄ, Y. 1971. "Constructing Sets of Uniformly Tighter Linear approximations for a Chance Constraint," Mgmt. Sci. 17:736-749.

SEPPÄLÄ, Y. 1972. "A Chance-Constrained Programming Algorithm," BIT 12:376-399.

SYMONDS, G.H. 1967. "Deterministic Solutions for a Class of Chance-Constrained Programming Problems," Op. Res. 15:495-512.

VADJA, S. 1972. Probabilistic Programming, Academic Press, New York.

VAN DE PANNE, C. and W. POPP 1963. "Minimum Cost Cattle Feed under Probabilistic Protein Constraints," Mgmt. Sci. 9:405-430.

WILKS, S.S. 1963. Mathematical Statistics, John Wiley, New York.

APPENDIX

Notations

Prob(.): Probability of the event argument.

E[.]: Expectation function.

V[.]: Variance function.

F(u;v): Cumulative density function of any random variable u evaluated at v, and $F^{-1}(u;.)$ is its inverse function.

n: Number of decision variables in the chance constraint.

a_i: Random technology coefficients with finite mean $E[a_i]$ and finite variance σ_i^2, i=1,..,n.

b: Resource parameter.

x_i: Decision variables, i=1,..,n.

α: Service performance target; probability target for satisfying the chance constraint. (Typically, $\alpha \in [0,1]$ is close to 1.)

$\phi(.)$: $\phi(s_1,..,s_n) = F^{-1}(\Sigma^n_{i=1} a_i s_i; \alpha)$ where $s_i \geq 0$, i=1,..,n.

z_α: Safety factor with service target α.

Z_α: Safety factor for normal distributions: one-tail normal variate for α.

RELATIVE ERROR'S
Averages for the Normal distribution

n	Max.-COV	HILL	SPL3	SPL6	OLSW	RAY1	RAY2	RAY3
2	0.02	1.43	0.22	0.21	0.51	0.51	0.08	0.08
	0.06	0.75	1.19	1.10	1.56	1.56	0.29	0.29
	0.20	3.57	2.22	1.57	3.17	3.17	0.80	0.80
	0.63	5.23	5.58	3.17	9.71	9.71	1.68	1.68
	2.00	17.66	3.49	0.96	19.44	19.44	2.98	2.98
4	0.02	0.18	0.53	0.52	0.89	0.89	0.51	0.11
	0.06	0.64	2.09	1.93	2.63	2.63	1.76	0.37
	0.20	2.55	5.37	4.19	7.20	7.20	4.44	0.99
	0.63	6.67	9.06	3.80	20.55	20.55	11.65	2.70
	2.00	13.85	5.86	1.37	36.75	36.75	22.06	5.72
6	0.02	0.21	0.67	0.64	1.08	1.08	0.80	0.11
	0.06	0.76	2.39	2.16	3.04	3.04	2.18	0.38
	0.20	1.93	7.51	5.67	10.21	10.21	7.12	1.11
	0.63	3.13	12.00	5.39	25.98	25.98	16.27	2.48
	2.00	9.44	5.33	1.42	60.96	60.96	45.72	6.34

RELATIVE ERROR'S
Standard Deviations for the Normal distribution

n	Max.-COV	HILL	SPL3	SPL6	OLSW	RAY1	RAY2	RAY3
2	0.02	3.76	0.21	0.21	0.22	0.22	0.05	0.05
	0.06	0.97	0.64	0.60	0.67	0.67	0.17	0.17
	0.20	6.45	1.47	1.22	1.82	1.82	0.40	0.40
	0.63	10.93	3.41	2.03	4.87	4.87	0.95	0.95
	2.00	27.43	2.16	0.90	7.38	7.38	1.42	1.42
4	0.02	0.25	0.27	0.26	0.31	0.31	0.28	0.07
	0.06	0.58	0.91	0.80	1.02	1.02	1.00	0.17
	0.20	3.09	2.36	1.71	3.23	3.23	2.34	0.45
	0.63	6.49	2.69	1.95	6.68	6.68	5.81	0.99
	2.00	11.72	4.25	1.09	10.07	10.07	11.00	2.10
6	0.02	0.23	0.44	0.42	0.50	0.50	0.37	0.07
	0.06	1.00	1.20	1.07	1.34	1.34	0.92	0.22
	0.20	1.55	2.57	1.73	3.69	3.69	2.78	0.58
	0.63	4.13	3.49	1.96	10.76	10.74	6.49	0.97
	2.00	7.93	3.46	1.95	14.55	14.55	15.79	2.57

Figure 2a. Results of Tests under Normal Distribution.

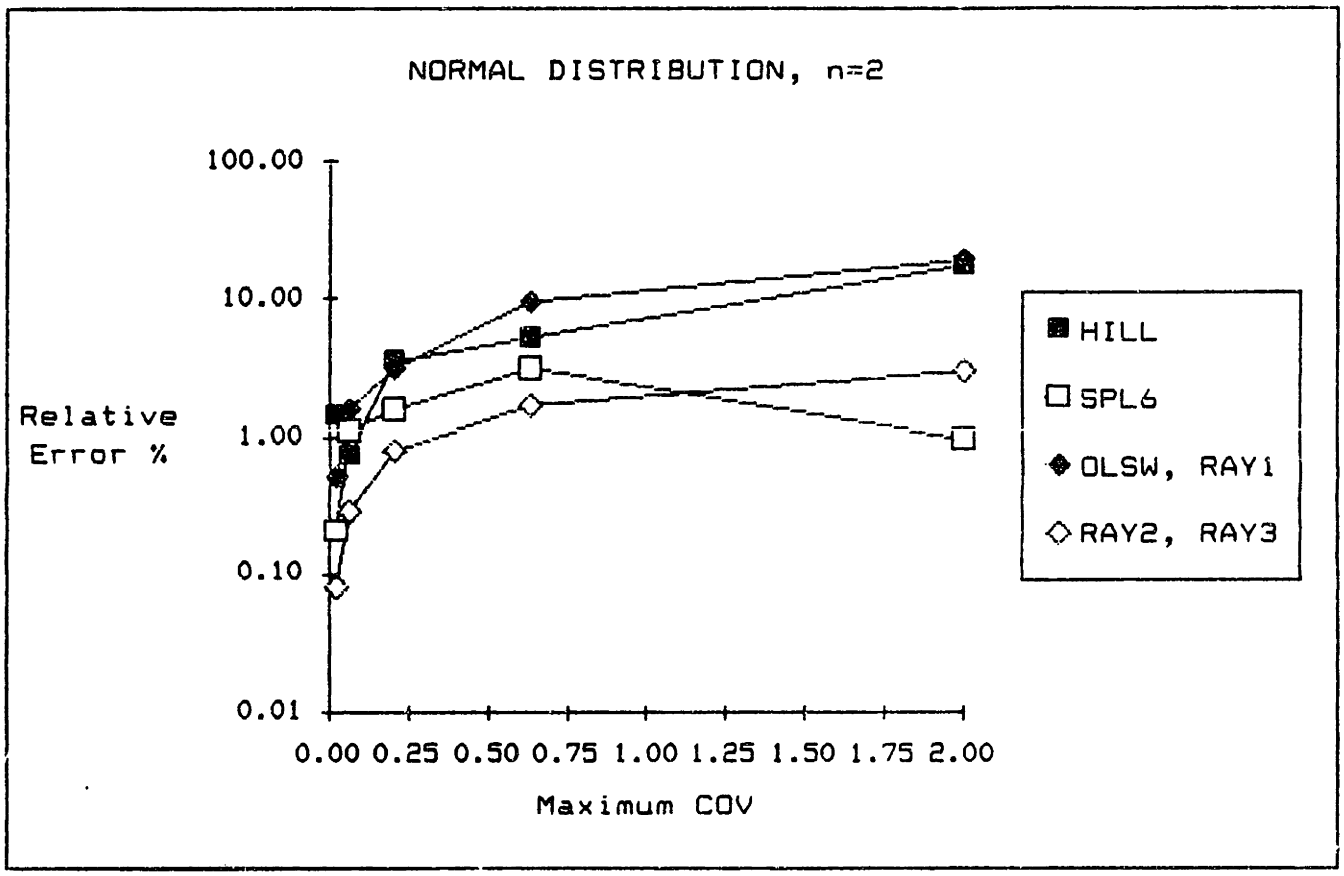


Figure 2b. Results of Tests under Normal Distribution.

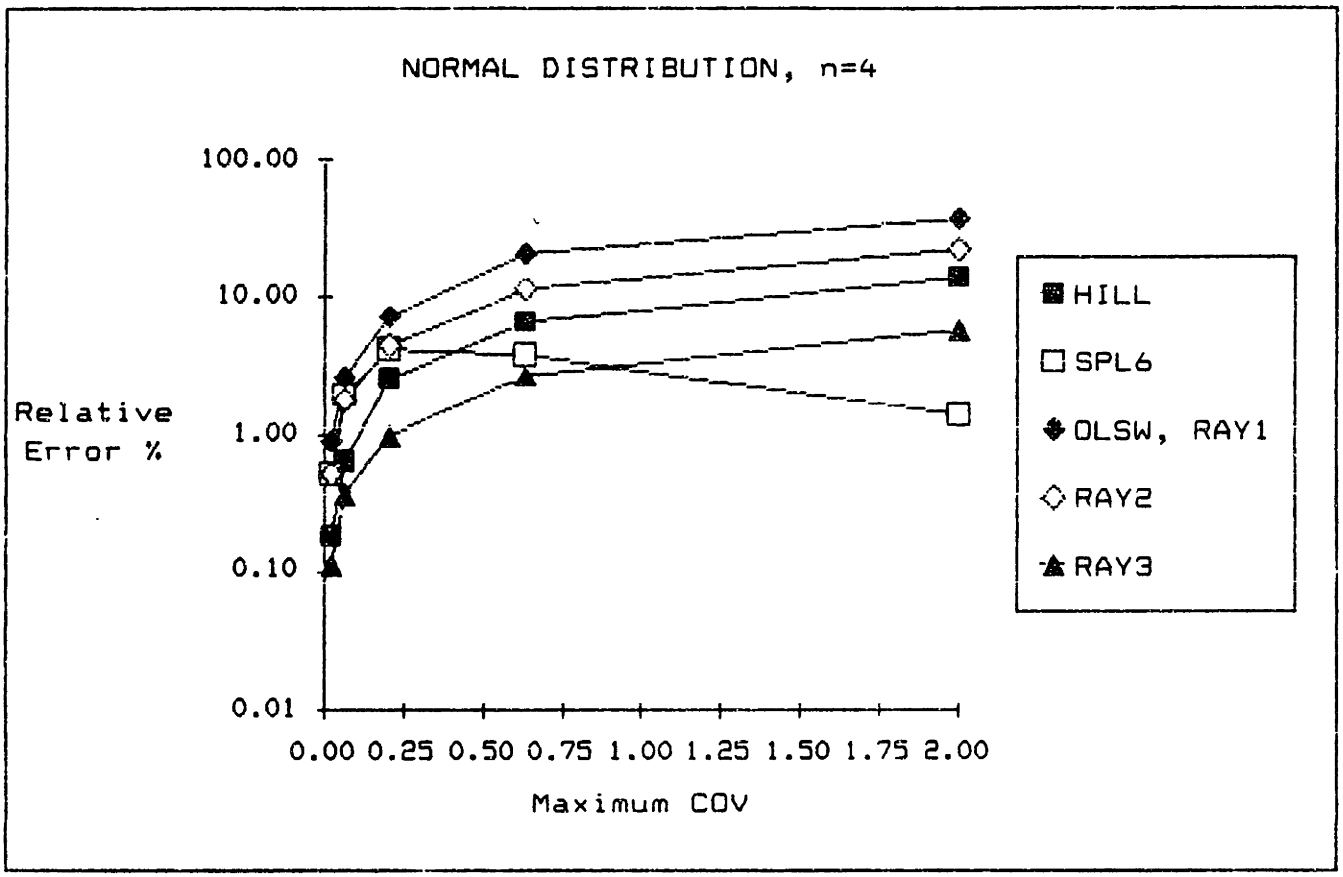


Figure 2c. Results of Tests under Normal Distribution.

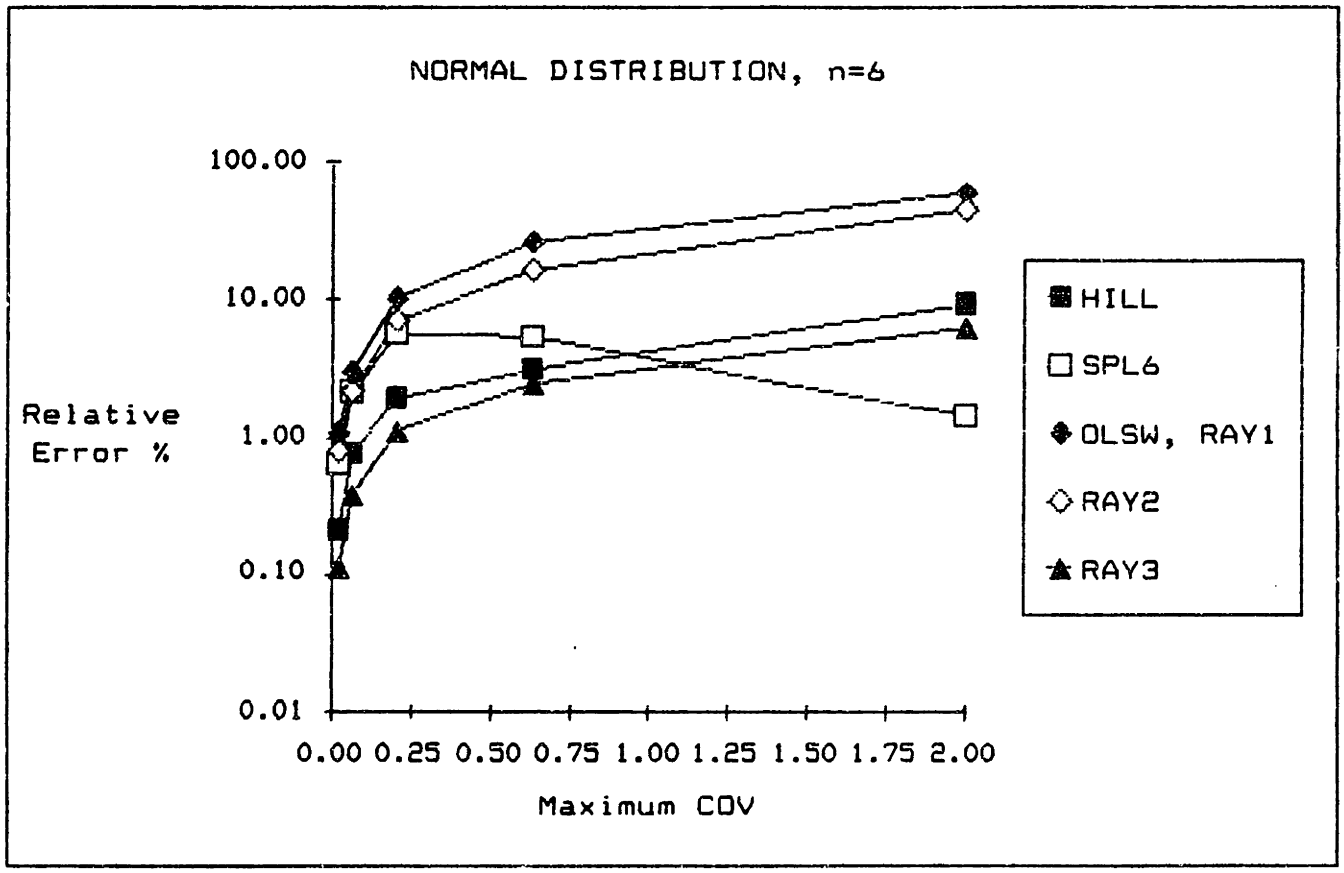


Figure 2d. Results of Tests under Normal Distribution.

RELATIVE ERROR'S
Averages for the Uniform distribution

n	HILL	SPL3	SPL6	OLSW	RAY1	RAY2	RAY3
2	79.75	85.30	81.20	91.42	3.42	0.00	0.00
4	81.41	87.85	75.54	106.75	11.70	7.88	0.00
6	71.89	92.30	72.90	121.80	19.84	17.42	0.00
8	63.55	94.86	70.50	131.94	25.32	25.32	----
16	44.46	86.65	52.26	158.95	39.91	39.91	----
32	28.49	72.46	36.13	171.52	46.70	46.70	----

RELATIVE ERROR'S
Standard Deviations for the Uniform distribution

n	HILL	SPL3	SPL6	OLSW	RAY1	RAY2	RAY3
2	9.16	0.91	6.29	10.88	5.88	0.00	0.00
4	25.63	3.28	9.83	20.63	11.14	9.33	0.00
6	27.21	5.67	10.48	18.23	9.85	9.98	0.00
8	12.23	3.78	8.06	12.06	6.52	6.52	----
16	11.82	11.05	13.10	8.47	4.58	4.58	----
32	5.09	8.56	7.54	1.97	1.06	1.06	----

Figure 3a. Results of Tests under Uniform Distribution.

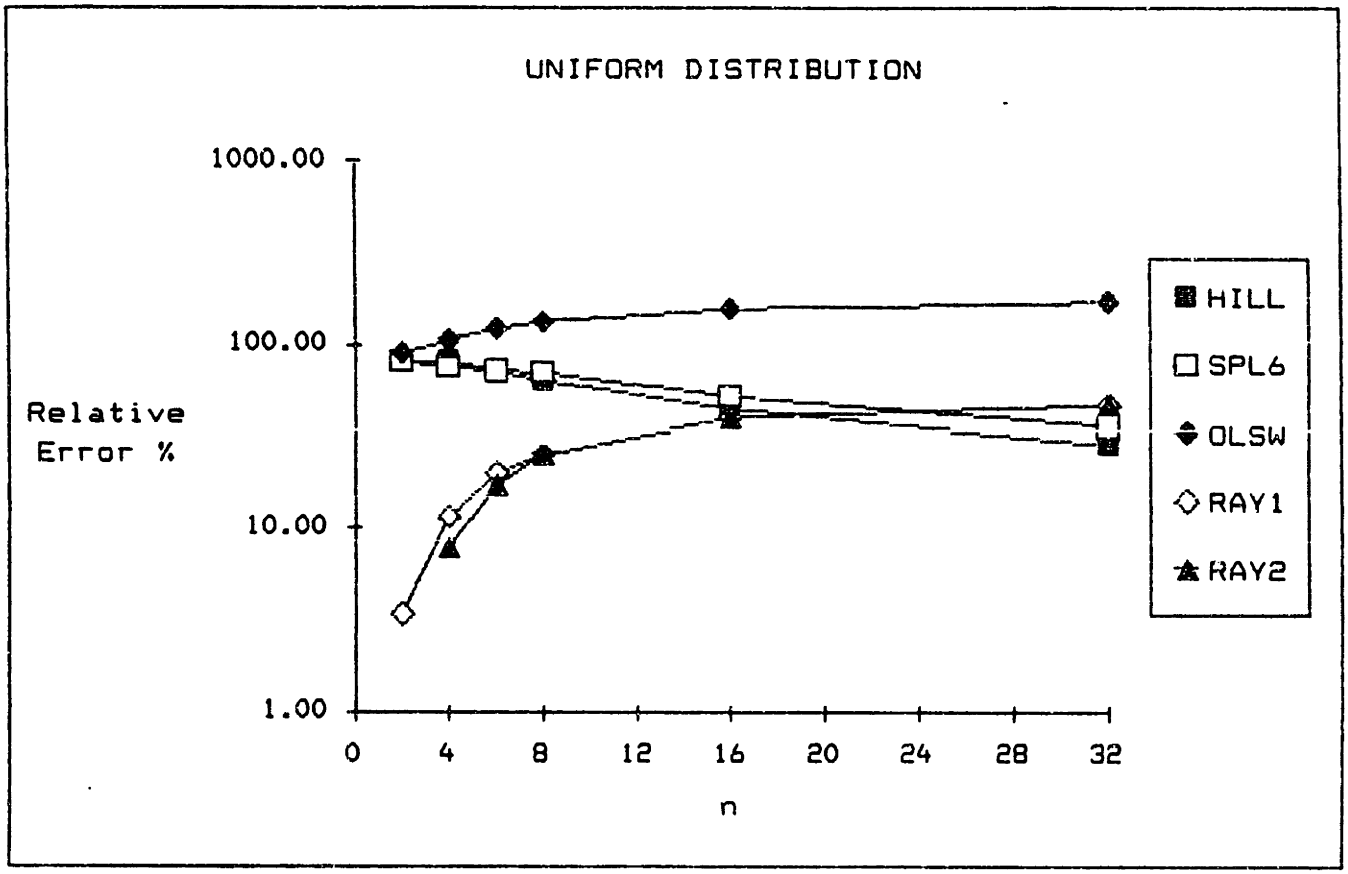


Figure 3b. Results of Tests under Uniform Distribution.

CHAPTER 6

CONCLUSION

1. Summary

The thesis examined three production planning problems with stochastic yield and substitutable demand. These problems are quite common but may have been neglected due to ignorance of the impact of stochasticity and the potential of substitutability. Also, problems with stochastic technology coefficients are difficult to solve. Hence, studies in this area may have also been limited by the level of analytic and computation technology available. Since computation technology have been making leaps of improvement, we hope that the analysis in this thesis took a step in matching it.

The problems examined share a set of core characteristics. This permitted us to study them in close unison. Each problem is interesting in its own right. But the various other conditions in the problems complicate the basic problem and require additional analysis. By examining the three problems, we studied and provided solutions to the stochastic yield and substitutable demand problems under a wide variety of accompanying conditions. We believe that there are other similar applications, especially in the provision of services, that can benefit from our analytical framework.

In this thesis, we first explored a way of structuring the interchangeability of products--referred to here as product demand substitution. We examined problems where the substitution is transitive and exploited that property to decouple joint chance constraints into

individual constraints. We also showed how non-transitive substitution problems can be transformed into transitive problems. In this way, general product demand substitution problems can be solved using the methods for transitive problems.

Unlike most stochastic yield problems, we seek the production plans (not control rules) for problems with service and capacity constraints. The uniformly-tighter formulations we provide linearize the original nonlinear problem. This permit us to use standard LP codes to solve these problems and make available the wealth of duality theory to assist further analysis.

2. Future Research

Additional work can be pursued in two directions: methodology and applications. There is still tremendous possibilities for improving on the method. For future research, we hope to consider the following: (a) examine how the relative errors can be parametrically bounded, (b) provide a multi-pass ε-optimal iterative approach, (c) consider problems with joint chance constraints, (d) provide theoretical results on the convexity of the feasible region.

We also desire to seek out and apply the method in more applications: airline reservations, investment management and other service problems, distribution problems with transhipment, part commonality manufacturing problems, assembly problems with stochastic yield, problems with demand for subassemblies, and large (stochastic or deterministic) nonlinear programming problems.