MANAGEMENT OF CO-PRODUCTION PROCESSES WITH RANDOM YIELDS: APPLICATIONS IN MANUFACTURING AND SERVICES

by

STEPHEN M. GILBERT

B.S.I.O.E. The University of Michigan (1984)
M.S.I.E. Stanford University (1985)

Submitted in Partial Fulfillment of the
Requirements for the Degree of

DOCTOR OF PHILOSOPHY
IN MANAGEMENT

at the

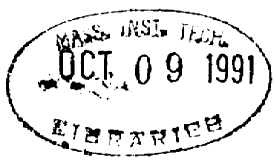
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

FEBRUARY 1992

Signature of Author_____
Sloan School of Management
August 5, 1991

Certified by _____
Gabriel R. Bitran
Thesis Supervisor

Accepted by_____
James B. Orlin
Chairman, Ph.D. Committee



ABSTRACT

A wide variety of operations management problems can be modeled as co-production with substitutable demand. That is, there are many situations in which the availability of two or more items are related, and because of randomness in either supply or demand, it can be advantageous to substitute one of these items for another.

For example, in the semi-conductor industry, chips are produced in large batches. Because of the presence of randomness in the process, individual chips in a given batch can perform differently. Because some customers have stricter specifications than others, chips are classified and sold as different products according to their measurable performance. In each period, the manufacturer faces a two-stage decision problem. First he determines the size of his production batch. Then, after the process is run, and the output is classified into product categories, he allocates chips to customers. Often this allocation decision involves trading-off the cost of backordering against the opportunity cost of substituting higher performance chips than the customer either needs or is willing to pay for.

This production and inventory problem is modeled as a dynamic program. Upper and lower bounds are developed for the cost of an optimal solution. Comparisons are made between these bounds and solutions that are obtained through heuristics.

The class of models studied in this dissertation can be applied to a wide variety of practical problems in both manufacturing and service industries. For example, a flexible machine can, in many cases, be thought of as a substitutable resource. Consider an expensive CNC flexible machine that performs both highly complex operations as well as some very simple operations that are ordinarily performed by less expensive machines. The production manager's decision to "substitute" the flexible machine for a less expensive one is analogous to the chip manufacturer's decision to substitute high performance chips for low ones.

Another example is that of a hotel reservation system. In many cases, hotels will offer various types of rooms and services for which they charge different rates. Customers may be willing to pay extra for a luxury suite, or the privilege of securing a room on short notice. A hotel manager must often decide whether to substitute a more luxurious room than a customer is willing to pay for or to turn him away.

The objective of this dissertation is to study these models and develop new insight and solution methods for a number of practical problems in the manufacturing and service industries.

Acknowledgements

I would like to express my sincerest gratitude to my advisor and friend, Gabriel R. Bitran, who has helped me to develop both professionally and personally. I will miss him very much when I depart from M.I.T.

I am grateful to Anne who has patiently endured many unsolicited lectures on random yields. The confidence and support that she has given me have been invaluable.

I am grateful to my parents, who have always encouraged me in the pursuit of my personal and professional goals.

Finally, I am grateful to my fellow students at MIT including: Dave Gebala, Sanjay Ghemawat, Suguna Pappu, Mike Peterson, Rob Smith, and Jim Walton. They have helped to make my stay here a pleasant one.

Table of Contents

PART ONE: Co-Production in the Semi-Conductor Industry

Chapter 1: Model Development

In the semi-conductor industry, the market for integrated circuits can be broken into two segments: specialty and commodity. As the name implies, specialty chips are custom designed and produced for a particular application. Because these chips are produced in low volumes, the fixed costs associated with designing the chips and procuring any specialized equipment that is necessary to produce them dominate the variable costs of production. As a result of this cost structure, customers often award a single manufacturer an exlusive contract to supply them with a particular type of specialty chip.

The process of selecting a manufacturer begins when a customer announces a "request for proposal" (RFP). An RFP typically includes an estimate of the volume of chips needed, the date by which they must be delivered, and a set of technical specifications. The level of detail for these specifications can range from a set of input and output parameters to a complete design. A manufacturer can respond to the RFP by preparing a proposal for producing the chips. A typical proposal might include a high level design, a discussion of the technical difficulties involved in producing the chips, and a price for which the manufacturer would be willing to produce them. The customer then evaluates the proposals from the various manufacturers on the basis of: price, the technical sophistication of the design, and the reputation of the manufacturer to deliver a quality product on schedule.

The market for commodity chips is much different. Manufacturers produce general purpose integrated circuits which can be used by a variety of customers. Because customers purchase "off the shelf", the transaction beween buyer and seller is much simpler than in the specialized segment. Here the customer selects an integrated circuit product which matches his technical requirements from among those that are available. Because commodity chips are frequently used as components in high volume consumer products, customers frequently require that delivery match their own production schedule. They often solicit bids from several different manufacturers to contract to supply a certain volume of chips per week or month at a pre-specified price. Contracts are awarded on the basis of price and the manufacturer's reputation for delivering a consistent product on time. A manufacturer's ability to succeed in this segment depends stongly upon how well he is able to reduce the variable costs of production while maintaining a consistent product quality and level of service.

The processes which manufacturers use to produce commodity integrated circuits can often be characterized as "co-production." As shown in Figure 1, a co-production process is one in which a family of several different products are produced simultaneously.

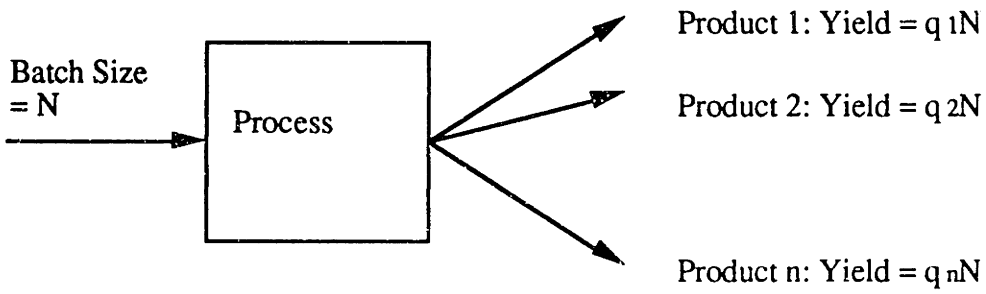


Figure 1: Co-production.

Although these products perform the same basic function, they differ from one another according to one or more key performance parameters. For example, suppose that several different customers need diodes, but each requires slightly different electrical performance. When a manufacturer produces a batch of diodes, the individual units may exhibit electrical properties which span a certain range. Thus, they can be classified as different products based upon where they fall within this range.

The first step in the production of semi-conductor chips is the drawing of ingots of either Silicon or Gallium Arsenide. These ingots are sliced into wafers. After several layers of semi-conducting material are placed on the wafers, they are cut into individual chips. Depending upon the complexity of the circuits involved, each wafer may yield between 10 and 100,000 chips. The individual chips can then be measured against one or more dimmensions of electrical performance and classified as products. A more detailed description of the production process can be found in Kothari(84) or Bitran and Tirupati (88).

The electrical performance of semi-conductor chips is extremely sensitive to changes in temperature, vibration, and the presence of dust during the manufacturing process. Because manufacturers cannot control these variables to the extent that they would like to, there can be a good deal of uncertainty as to the electrical properties that will be exhibited by the chips in a single production run. For example, suppose that a diode can be classified as one of three different products (1, 2, and 3) on the basis of two measureable electrical properties (Property A and Property B). A possible classification scheme is shown in Figure 2. In this figure, Properties A and B are respresented on the X and Y axis of a graph. Diodes could be

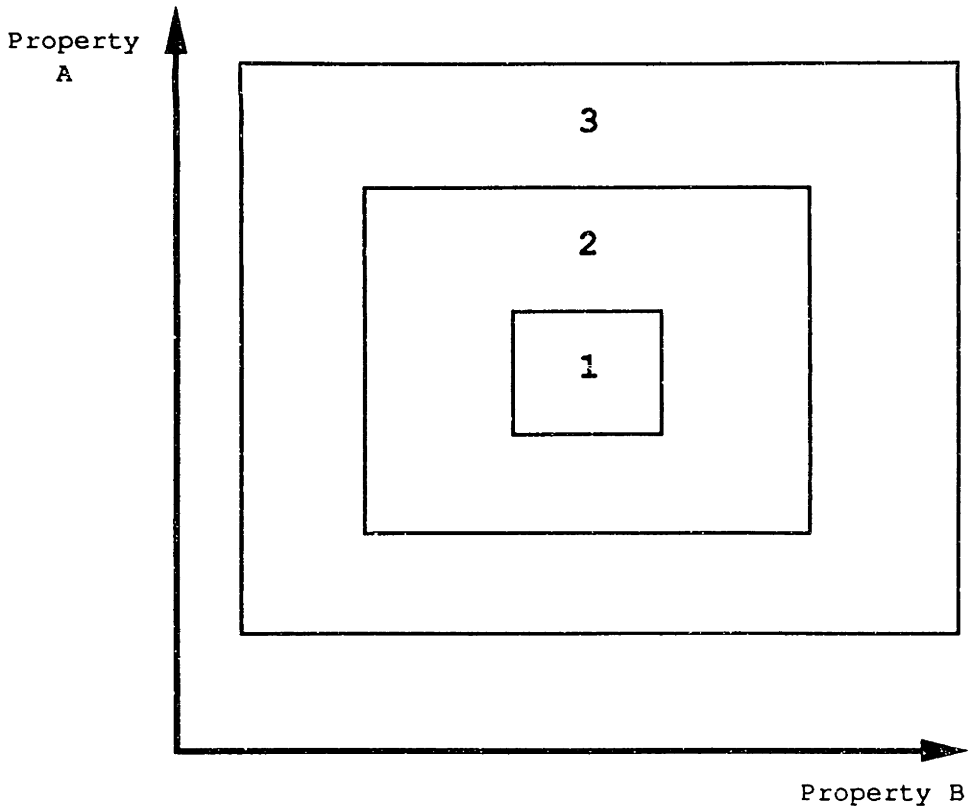


Figure 2: Serially Nested Product Structure

classified as 1, 2, or 3 according to the zone into which their properties map. A single production batch may contain all three diode products, and the fractions that can be classified as products A, B, and C are often random variables.

In the classification scheme that is depicted in Figure 2, the specifications for a given product are strictly looser than those for products with lower indices, and strictly tighter than those for higher indices. Notice that zone 1 is completely contained within zone 2, and zone 2 is similarly contained within zone 3. This "serially nested" product structure implies that the demand for a given product i can be satisfied with products which fall into categories i or lower. For example, a customer who has product 3 would be equally happy to receive any of the three products; A

customer who has ordered product 2 would be satisfied with either product 2 or 1; A customer who has ordered product 1 can be satisfied only with the delivery of that specific product.

Such product structures are common in practice. Often, different customers need the same basic product, but some need tighter tolerances than others. One example is found in memory chips. Different customers need different size chips. If there is a defect in a single quadrant of a large chip, it may not be necessary to scrap the entire unit. The remaining good quadrants can be separated and supplied as smaller chips. Downgrading is possible by separating the quadrants of a non-defective large chip.

Another example is found in the production of diodes. These devices function as electronic valves. When a batch of diodes is produced, individual units in the batch may demonstrate different levels of resistance on forward or negative bias. Because some customers are willing to accept higher levels of resistance than are others, it is possible to classify products according to their measureable performance. Customers with the strictest specifications accept only those diodes with the lowest levels of resistance. Here, downgrading is defined as providing a customer with a diode which satisfies a lower maximum tolerance for resistance than he needs.

Because semi-conductor chips are often supplied in accordance to long term contractual obligations, demand is known far in advance. The production manager's job is to minimize the costs of production, inventory, and backorders. Because random effects influence the relative quantity of each type of chip in a given batch, two decisions are made in each period. At the beginning of the period, in the "morning", the production manager makes a lot sizing decision: how many total chips to produce. After the

batch is produced, he observes the number of chips which fall into each product classification.

At this point he faces the "afternoon problem": how to allocate chips to customers. This problem is faced when the yield of a given chip type i exceeds its demand while, at the same time, there is an insufficient supply of a less strictly specified (higher indexed) chip type j. In general, the penalties for backorders are highest for the most strictly specified products. Thus it would not be advantageous to downgrade a given chip type unless its own demand was satisfied first. By downgrading product i to satisfy demand for j, the backorder and holding costs in the current period are reduced. However, since the production quantity decision for the following period depends on the inventory levels, the allocation of products to customers affects the costs in future periods.

Before formulating an analytical model of this problem, let us define the following notation:

N_t : The production quantity in period t.

W_{it}^- : The amount downgraded from product i to i+1 in period t.

$\mathbf{W}_t^-$: A vector whose components are $W_{1t}^-,\ldots, W_{nt}^-$.

W_{it}^+ : The amount downgraded from product i-1 to i in period t.

$\mathbf{W}_t^+$: A vector whose components are $W_{1t}^+,\ldots, W_{nt}^+$.

J_{it} : The inventory of product i at the end of period t.

J_{it}^+ : $\text{Max}\{0, J_{it}\}$

J_{it}^- : $\text{Min}\{0, J_{it}\}$

$\mathbf{J}_t$: A vector whose components are $J_{1t},\ldots, J_{nt}$.

d_{it} : Demand for product i in period t.

$\mathbf{d}_t$: A vector whose components are $d_{1t},\ldots,d_{nt}$.

q_{it} : The yield of product i in period t. (The fraction of the production quantity in period t which meets the specifications for product i, but fails to meet the specifications for products i-1,...,1.)

$\mathbf{q}_t$: A vector whose components are $q_{1t},...,q_{nt}$.

h_i : Per period inventory holding cost for product i.

p_i : Per period backorder cost for product i.

r : Per period discount factor.

c : Production cost.

Note that the parameters of the problem consist of demand, the discount factor, and the costs of production, backorders, and inventory. The decision variables in each period are the lot size, and the quantity of each product to downgrade. Although the inventory levels are functions of these decisions and demand, they are also influenced by the effects of random yields.

The production manager's problem can be modeled as a nested dynammic program:

(S)

$$M_1(\mathbf{J}_0) = \min_{N_1} \left\{ cN_1 + \int_{\mathbf{q}_1} A_1(\mathbf{J}_0, N_1, \mathbf{q}_1) \cdot f(\mathbf{q}_1) \cdot d\mathbf{q}_1 \right\} \qquad 1.1$$

where:

$$A_t(\mathbf{J}_{t-1}, N_t, \mathbf{q}_t) = \min_{\mathbf{W}_t^+, \mathbf{W}_t^-} \left\{ \sum_{i=1}^{n} h_i J_{it}^+ + \sum_{i=1}^{n} p_i J_{it}^- + rM_{t+1}(\mathbf{J}_t) \right\}, \quad \text{for } t = 1,...,T-1 \qquad 1.2$$

$$M_t(\mathbf{J}_{t-1}) = \min_{N_t} \left\{ cN_t + \int_{\mathbf{q}_t} A_t(\mathbf{J}_{t-1}, N_t, \mathbf{q}_t) \cdot f(\mathbf{q}_t) \cdot d\mathbf{q}_t \right\} \qquad 1.3$$

$$A_T(\mathbf{J}_{T-1}, N_T, \mathbf{q}_T) = \min_{\mathbf{W}_T^+, \mathbf{W}_T^-} \left\{ \sum_{i=1}^{n} h_i J_{iT}^+ + \sum_{i=1}^{n} p_i J_{iT}^- \right\} \qquad 1.4$$

$$\mathbf{J}_{t-1} - \mathbf{J}_t + \mathbf{q}_t N_t + W_t^+ - W_t^- = \mathbf{d}_t, \quad \text{for } t = 1,\ldots,T \tag{1.5}$$

$$W_t^- \le \text{Max}(0, \mathbf{J}_{t-1} + \mathbf{q}_t N_t + W_t^+), \quad \text{for } t = 1,\ldots,T \tag{1.6}$$

$$N_t \ge 0, \quad \text{for } t = 1,\ldots,T \tag{1.7}$$

$$W_{it}^- = W_{i+1,t}^+, \quad \text{for } i = 1,\ldots,n-1 \text{ and } t = 1,\ldots,T \tag{1.8}$$

$$W_{it}^+ \ge 0, \; W_{it}^- \ge 0, \quad \text{for } i = 1,\ldots,n-1 \text{ and } t = 1,\ldots,T \tag{1.9}$$

$$W_{nt}^+ = W_{1t}^- = 0, \quad \text{for } t = 1,\ldots,T \tag{1.10}$$

In this formulation, the objective is to minimize the expected costs of production, inventory, and backorders over a T period horizon. The use of the expected value function is justified on the basis that the process is assumed to run for a large number of periods. Under most circumstances, it can safely be assumed that the production manager is risk neutral. Each period consists of a two stage dynamic program: $M_t(\mathbf{J}_{t-1})$ represents the lot sizing decision that will made at the beginning (in the "morning") of period t. $A_t(\mathbf{J}_{t-1}, N_t, \mathbf{q}_t)$ represents the downgrading decisions that will be made after the yields $\mathbf{q}_t$ are observed in period t (in the "afternoon"). $A_T(\mathbf{J}_{T-1}, N_T, \mathbf{q}_T)$ represents the downgrading decisions that are made in the final period of the planning horizon. In some cases, the costs for inventory and backorders are different in period T than in earlier ones. We have assumed that, although the planning horizon is T periods, the actual manufacturing process will continue far beyond T. Thus backorder and penalty costs in period T are the same as in the previous ones. Constraint 1.5 is an inventory balance constraint. Constraint 1.6 represents the fact that we can only downgrade an item to the extent that it is physically available. That is, it does not allow downgrading which would increase the backorders of a given product in the current period. In most cases, such downgrading would not occur anyway; penalty costs tend to be higher for the more tightly

specified products, so there is no advantage to backordering a product in order to satisfy demand for a more loosely specified one. Thus, because constraint 1.6 is unlikely to be binding in an optimal solution, we will exclude it from this point forward.

Analysis of model (S) is extremely difficult. The yields of the different products tend to be correlated with one another. Furthermore, because of the nesting of the two stage dynamic programs, the probability distributions for the inventory levels in periods 2,...,T are extremely complicated. However, by approximating and analyzing the model, we can gain insight to motivate heuristic solution procedures.

Although there is a relatively large literature on managing production systems with random yields, most of the work has focused on one-stage, one product environments. Mazzola, McCoy, and Wagner (1987) provide a good review of the Variable Yield Lot Sizing Problem (VYLP) literature. More recently, Yano and Lee (1989) provide a taxonomy of the literature in lotsizing with random yields. Several significant contributions have been made since these surveys. Taking a slightly different perspective on random yields, Singh, Abraham, and Akella (1990) describe a problem in which sites on a semi-conductor wafer are allocated to various types of chips in order to maximize the probability of completing a set of chips by some due date. Wein and Avram (1990) worked on a similar problem, but considered multiple set types and used throughput as the objective.

Lee (1990) also studies a problem that differs from most of those in the literature in that it is assumed that the process operates in one of two states: in or out of control. He decribes heuristics that perform well with respect to an enumerated search optimal solution.

Because models of multi-stage production systems with variable yields are inherently complex, this literature is much more limited. Tang (1990) provides an up to date review of the work that has been done.

Unfortunately, the problem which we address is relatively unresearched. Although co-production of substitutible products is a phenomena that can be observed in a variety of both high and low tech manufacturing environments, it has remained relatively unrecognized by the academic community. The first known reference to the problem is Bitran and Dasu (1988). Their model seeks to maximize expected profits in an environment with deterministic demand, and production, backorder, and inventory holding costs. Discrete yield distributions to approximate reality. However the size of the problem grows exponentially in both the number of possible yield outcomes and the length of the horizon. Bitran and Leong (1990) considered continuous random variables and modeled the problem as one of minimizing the expected costs of production and inventory subject to service level constraints. They construct linear approximations to these stochastic constraints, and develop heuristic procedures for solving the problem. Ou and Wein (1990) considered a closely related problem. In their model, several different processes are available. Although each one is targeted at producing one of a hierarchically ordered set of products, it produces lower level products with some probability. They develop heuristics on the basis of insights gained from brownian motion approximations of the scheduling problem. However, their model differs from ours in that they do not allow for the downgrading of products.

In the sections that follow, we continue the efforts of the aforementioned authors in analyzing the problem and developing

heuristics. In section 2, we show how the model can be approximated. In section 3, we use these approximations, the work of Bitran and Dasu (1988) and that of Bitran and Leong (1990), as well as the advice of production managers to derive heuristics for solving the model on a rolling horizon basis. Section 4 describes a theoretical lower bound on the value of an optimal solution. Although this bound cannot be calculated in closed form, we discuss a method for estimating it using Monte Carlo simulation. In Section 5, we compare the performance of our heuristics to the lower bounds by using Monte Carlo simulation. Finally, in Section 6, we summarize our findings and explore potential directions for future research.

Chapter 2: Model Analysis

Because the nested dynamic programs in M are so difficult to annalyze, it is necessary to approximate the original model. One approach is to model the problem as a static production planning problem with service level constraints. In this approach, the costs of backorders are removed from the objective function and replaced with service level constraints. An item's service level is defined as the probability that there is non-negative inventory of that item at the end of a period. The constraints require that each item i have a service level of at least α_i in each period, and α_i is chosen in such a manner to represent the trade-off between inventrory holding and backorder costs. Conceptually, this is similar to the single period newsboy model in which the optimal purchase quantity is chosen so that the probability of a stockout is equal to h/(h+p), where h and p are the per unit costs of inventory and shortage respectively.

We use this idea to create a production planning problem with service level constraints as a surrogate to our original stochastic dynamic program. By assuming that all production lotsizing and downgrading decisions are made at the beginning of the horizon, we are left with th following non-linear stochastic program:

(SC)

$$Z_{SC} = \text{Min } E\left(\sum_{i=1}^{n}\sum_{t=1}^{T} h_i J_{it}^{+} + c\sum_{t=1}^{T} N_t\right) \qquad 2.1$$

subject to:

$$J_{t-1} - J_t + q_t N_t + WS_t^{+}\left(\sum_{\tau=1}^{t} W_\tau^{+}\right) - WS_t^{-}\left(\sum_{\tau=1}^{t} W_\tau^{-}\right) = d_t, \quad \text{for } t= 1,...,T \qquad 2.2$$

$$\text{Prob}\,(J_{it} \geq 0) \geq \alpha_i, \qquad i = 1,...,n;\ t = 1,...,T \tag{2.3}$$

$$N_t, W^-_t, W^+_t \geq 0, \qquad i = 1,...,n-1;\ t = 1,...,T \tag{2.4}$$

where WS_t^+ and WS_t^- are n dimensional vector valued functions whose ith components are respectively :

$$WS^-_{it}\left(\sum_{\tau=1}^{t} W^-_\tau\right) = \text{Max}\left\{0,\ \text{Min}\left\{\sum_{\tau=1}^{t} W^-_{i\tau},\ \sum_{\tau=1}^{t}(q_{i,\tau}N_\tau - d_{i,\tau}) + WS^-_{i-1,t}\left(\sum_{\tau=1}^{t} W^-_{i-1,\tau}\right)\right\}\right\}, \text{ and}$$

$$WS^+_{it}\left(\sum_{\tau=1}^{t} W^+_\tau\right) = \text{Max}\left\{0,\ \text{Min}\left\{\sum_{\tau=1}^{t} W^+_{i\tau},\ \sum_{\tau=1}^{t}(q_{i,\tau}N_\tau - d_{i,\tau}) + WS^+_{i-1,t}\left(\sum_{\tau=1}^{t} W^+_{i-1,\tau}\right)\right\}\right\},$$

for i = 1,...,n-1, and t = 1,...,T.

In this optimization problem, the vector valued functions $WS_t(X)$ are random variables representing the total amount downgraded from product i between periods 1 and t. These functions play the same role in this model that constraint 1.6 plays in model (S). Recall that in most cases, because backorder penalties are higher for more strictly specified products, we can drop constraint 1.6 from model (S). However, because of the probabilistic constraints in problem (SC), it is less obvious that we can ignore the $WS_t(X)$ functions.

The ith component of the sum of the decision vectors $W_1^-,...,W_t^-$ represent the total amount that we would like to downgrade from i to i+1 in periods 1 through t. Similarly, the ith component of the sum of the decision vectors $W_1^+,...,W_t^+$ represents the total amount that we would like to downgrade from i-1 to i in periods 1 through t. Note that in order to facilitate vector notation, we have constructed vectors $W_1^-,...,W_t^-$ and $W_1^+,...,W_t^+$ where: $W^+_{1t} = W^-_{nt} = 0$, and $W^+_{it} = W^-_{i-1,t}$, for i = 2,...,n.

Because of the service level constraints (equation 2.3), we will have negative inventory of product i with a probability of up to $1-\alpha_i$. When the

inventory of a given item i is negative, we may be prevented from downgrading as much as we would like to into item i+1. The actual amount that is downgraded from i to i+1 by period t is a function of the decision variables W_{i1}^-,...,W_{it}^- and W_{i1}^+,...,W_{it}^+ as well as the availability of product i.

Unfortunately, the random downgrading functions make SC very difficult to analyze. We can approximate SC with:

(SCA):

$$Z_{SCA} = \text{Min } E\left(h\sum_{i=1}^{n}\sum_{t=1}^{T} J_{it}^{+} + c\sum_{t=1}^{T} N_t \right) \qquad 2.5$$

subject to:

$$J_{t-1} - J_t + q_t N_t + \sum_{\tau=1}^{t} W_\tau^+ - \sum_{\tau=1}^{t} W_\tau^- = d_t, \quad \text{for } t= 1,...,T \qquad 2.6$$

$$\text{Prob}\,(J_{it} \geq 0) \geq \beta_i, \qquad i = 1,...,n; \quad t = 1,...,T \qquad 2.7$$

$$\alpha_i \leq \beta_i \leq 1, \text{ and } \alpha_i \leq \prod_{j=1}^{i} \beta_j, \qquad i = 1,...,n \qquad 2.8$$

$$N_t, W_{it} \geq 0, \qquad i = 1,...,n-1; \quad t = 1,...,T \qquad 2.9$$

This formulation differs from (SC) only in that constraint 2.2 has been replaced with 2.6, and that constraint 2.3 has been replaced with constraints 2.7 and 2.8.

Claim: A solution that is feasible in (SCA) is also feasible in (SC), and its cost represents an upper bound on the cost of an optimal solution to (SC).

Proof: The proof is by induction on i. Suppose that (N,W) is a feasible solution to (SCA). Let us define J_t(SCA) to be the value of J_t as defined in constraint 2.6 of problem (SCA) given the solution in terms of (N,W).

Similarly let us define $\mathbf{J}_t(SC)$ to be the value of $\mathbf{J}_t$ as defined in constraint 2.2 of problem (SC) given the solution in terms of (N,W).

Because $W_{1t} \geq WS_{1t}$ $(W_{1t}) \geq 0$, whenever J_{1t} (SCA) is non-negative, $J_{1t}(SC)$ will be non-negative. Thus, because $\beta_1 \geq \alpha_1$, constraint 2.7 implies 2.3 for product 1.

Now suppose that, for all t = 1,...,T, the solution (N,W) satisfies the service level constraints in problem (SC) for product i. That is:

$$\text{Prob}(J_{it}(SC) \geq 0) = \text{Prob}\left[\sum_{\tau=1}^{t} (q_{i,\tau}N_\tau - d_\tau) + WS_{i-1,t}\left(\sum_{\tau=1}^{t} W_{i-1,\tau}\right) - \sum_{\tau=1}^{t} W_{i,\tau} \geq 0\right] \geq \alpha_i.$$

Whenever $J_{it}(SC)$ is non-negative, we will be able to downgrade as much as we would like from i to i+1. Thus:

$$\text{Prob}\left(WS_{i,t}\left(\sum_{\tau=1}^{t} W_{i,\tau}\right) = \sum_{\tau=1}^{t} W_{i,\tau}\right) = \text{Prob}[J_{it}(SC) \geq 0] \geq \alpha_i.$$

The relationship between constraints 2.7 of (SCA) and constraints 2.3 of (SC) can be expressed in the following conditional probability statement:

$$\text{Prob}\left(J_{i+1,t}(SC) \geq 0 \mid WS_{i,t}\left(\sum_{\tau=1}^{t} W_{i,\tau}\right) = \sum_{\tau=1}^{t} W_{i,\tau}\right) \geq \beta_{i+1}.$$

Or, equivalently:

$$\text{Prob}\left(J_{i+1,t}(SC) \geq 0 \mid J_{i,t}(SC) \geq 0\right) \geq \beta_{i+1}.$$

In words, this says that, given that there is sufficient inventory of product i to downgrade as much as we would like to, the likelihood of non-negative inventory of product i+1 is at least β_{i+1}. Thus, the likelihood that $J_{i+1,t}$, as defined in equations 2.2 of problem SC, is non-negative is:

$$\text{Prob}[J_{i+1,t}(SC) \geq 0] = \text{Prob}\left[J_{i+1,t}(SC) \geq 0 \mid J_{i,t}(SC) \geq 0\right] \cdot \text{Prob}[J_{i,t}(SC) \geq 0]$$

$$\geq \beta_{i+1} \cdot \alpha_i \geq \frac{\alpha_{i+1}}{\prod_{j=1}^{i} \beta_j} \alpha_i \geq \frac{\alpha_{i+1}}{\alpha_i} \alpha_i = \alpha_{i+1}.$$

where the last two inequalities follow from constraint 2.8 of (SCA).◊

The implication of the above result is that an optimal solution to the approximation (SCA) will provide an upper bound on the cost of an optimal solution to problem (SC). Because of the nonlinearity of the convex service level constraints (equations 2.7), problem (SCA) is difficult to solve. However, it is possible to derive insight from its formulation, and by using interior linear approximations of the convex service level constraints (equations 2.7), it is possible to obtain a bound on the value of its optimal solution. It can be shown that, by increasing the number of linear constraints, this bound can be arbitrarily close to the optimal cost.

In order to gain some insight into the the original co-production problem, it helps to express the service level constraints (2.7) of problem (SCA) in an equivalent deterministic form. This deterministic equivalent is based on the following well known result:

Claim: For a continuous random variable X from a known distribution with non-zero, finite variance σ_X, and any real numbers N, Y and α on the open interval (0,1), there exists a real value K_α such that: If $N(E[X] + K_\alpha \sigma_X) \geq Y$, then Prob $[Nx \geq Y] \geq \alpha$.

For the interested reader, a proof is given in Symonds (1967). We can use the above result and the fact that:

$$E\left[\sum_{\tau=1}^{t} N_\tau q_{i\tau}\right] = E[q_i]\sum_{\tau=1}^{t} N_\tau, \text{ and } Var\left[\sum_{\tau=1}^{t} N_\tau q_{i\tau}\right] = \sum_{\tau=1}^{t} N_\tau^2 \sigma_i^2,$$

to re-write (SCA) in an equivalent deterministic form. Let us first define:

$\bar{q}_i$: The expected value of the yield for product i, assumed to be time invariant.

$\bar{\mathbf{q}}$: A vector who's elements are $\bar{q}_1, \ldots, \bar{q}_n$.

σ_i : The standard deviation of the yield for product i, assumed to be time invariant.

σ : A vector whose elements are $\sigma_1,\ldots,\sigma_n$.

K_i : The number of standard deviates from the mean of the distribution for q_i that corresponds to the fractile α_i as shown above.

$\mathbf{K}$: An n x n diagonal matrix in which the ith element along the diagonal is equal to K_i.

Using these definitions, we the deterministic equivalent to (SCA) is as follows:

(SCAD):

$$Z_{SCAD} = \text{Min } E\left(\sum_{i=1}^{n}\sum_{t=1}^{T} h_i J_{it}^{+} + c\sum_{t=1}^{T} N_t\right) \qquad 2.10$$

subject to:

$$\mathbf{J}_{t-1} - \mathbf{J}_t + \mathbf{q}_t N_t + \sum_{\tau=1}^{t} \mathbf{W}_\tau^{+} - \sum_{\tau=1}^{t} \mathbf{W}_\tau^{-} = \mathbf{d}_t, \quad \text{for } t= 1,\ldots,T \qquad 2.11$$

$$\sum_{\tau=1}^{t} N_\tau \cdot \bar{\mathbf{q}} - \sqrt{\sum_{\tau=1}^{t} N_\tau^2} \cdot \mathbf{K}\,\sigma + \sum_{\tau=1}^{t} [\mathbf{W}_\tau^{+} - \mathbf{W}_\tau^{-}] \geq \sum_{\tau=1}^{t} \mathbf{d}_\tau, \quad t = 1,\ldots,T \qquad 2.12$$

$$N_t, \mathbf{W}^{-}_t, \mathbf{W}^{+}_t \geq 0, \qquad i = 1,\ldots,n-1;\ t = 1,\ldots,T \qquad 2.13$$

Where $\bar{q}$ is defined as an n dimensional vector such that the i^{th} element is equal to $E[q_i]$. Consider this problem under the following conditions:

$$d_{it} = d_i,\ E[q_{it}] = E[q_i] \quad \forall\ i = 1, \ldots, n \tag{2.14}$$

$$\frac{d_i}{E[q_i]} = R \tag{2.15}$$

$$K_1 cv_1 = K_1 \frac{\sigma_1}{E[q_1]} = \max_i \left\{ K_i cv_i = K_i \frac{\sigma_i}{E[q_i]} \right\} \tag{2.16}$$

Condition 2.14 requires that the demand and expected yield be constant for each product. Because chips are often supplied according to long term contractual obligations, the industry tends to be cyclical with respect to demand. But the cycles are generally long enough that demand can be treated as though it were constant for production planning purposes. Technological breakthroughs can also improve expected yields. But in between these breakthroughs, yields are very stable. Thus condition 2.14 is often valid in practice.

Condition 2.15 says that the ratio of demand to expected yield is constant across products. Although this may not always hold, the production planning problem is the most difficult, and the most interesting, when it does. Consider, for example, the product hierarchy shown in Figure 2. If the ratio of demand to expected yield for product 1 were considerably higher than for the lower grade products, then the lower grade products could be considered to be bi-products. Product 1 would drive lot sizing decisions, and there would tend to be more than enough production of the lower level products to satisfy demand.

The other case, where the ratio of demand to expected yield is considerably lower for the highest grade than for lower grades is not often found in practice. The highest grades typically command premium prices, and manufacturers are particularly eager to supply them. However, there are usually large costs involved in modifying the process to obtain better yields of these products. Thus, it is most often the case that contractual obligations, i.e. demand, is at least as high relative to the yields of the process for the high grade products as it is for the lower grades.

Condition 2.16 concerns the safety margins that are required to meet the service level constraints for the various products. In order to satisfy the service level constraints for a given product, it is necessary that the expected production yield exceed demand by a certain safety factor. When the yield distribution is known, this safety factor can be determined as a function of the required service level for the product, and the variance of the yield distribution. Condition 2.16 says that the the ratio of this safety margin to the expected yield be at least as large for product 1 as for any other product. In practice, it is often the case that, the service level requirements are at least as high, if not higher, for the best grade products than for the others. Because the coefficient of variation for the yield of this product may also be high, Condition 2.16 is valid in a large number of practical cases.

Claim: Suppose that conditions 2.14, 2.15, and 2.16 hold, and that the yield distributions of products i = 1,...,n share the same stable distributional form. (See Allen, Braswell, and Rao (1974) for a discussion of the characteristics of stable probability distributions.). If $h_i \leq h_{i+1}$, and $cv_i >$

cv_{i+1}, for i = 1,...,n-1 , then in any optimal solution to SCAD , $W_{i\tau} = 0$, for all i = 1,...,n-1, and τ = 1,...,T.

Proof: Because no other product downgrades to product 1, the service level constraints 2.11 for product 1 imply that every feasible solution vector N = $(N_1, ..., N_T)$ must satisfy:

$$\sum_{\tau=1}^{t} N_\tau \geq \frac{td_1 + W_1 + K_1\sigma_1\sqrt{\sum_{\tau=1}^{t} N_\tau^2}}{E[q_1]} \geq \frac{td_1 + K_1\sigma_1\sqrt{\sum_{\tau=1}^{t} N_\tau^2}}{E[q_1]}, \quad t = 1,..., T.$$

Because $K_1cv_1 \geq K_icv_i$, for i = 2,...,n, we can now use condition 2.15 to show that any N = $(N_1, ..., N_T)$ which satisfies the service level constraints for product 1 automatically satisfies the corresponding constraints for the other products when the vectors W^-_t and W^+_t are forced to zero, i.e. downgrading is not allowed. We have:

$$\sum_{\tau=1}^{t} N_\tau \geq \frac{td_1 + K_1\sigma_1\sqrt{\sum_{\tau=1}^{t} N_\tau^2}}{E[q_1]} = tR + K_1\, cv_1 \geq tR + K_i\, cv_i = \frac{td_i + K_i\sigma_i\sqrt{\sum_{\tau=1}^{t} N_\tau^2}}{E[q_i]}$$

for all i= 1,...,n.

Because the objective function penalizes ΣN_τ, the service level constraint for product 1 in period T will be binding in any optimal solution. Since downgrading cannot reduce the total amount of production that will be required to satisfy this final service level constraint for product 1, downgrading will only be done if it reduces the expected holding costs.

Let $N = (N_1, ..., N_T)$ be a vector that satisfies the service level constraints for product 1, and consider the sum of the expected holding costs for products i = 1,...,n in period t:

$$\sum_{i=1}^{n} h_i \int_{x=td_i-\sum_{\tau=1}^{t} W^{+}_{i\tau}+\sum_{\tau=1}^{t} W_{i\tau}}^{\infty} \left(x - td_i - \sum_{\tau=1}^{t} W^{+}_{i\tau}+\sum_{\tau=1}^{t} W_{i\tau}\right) f_i(x)dx \qquad 2.17$$

Note that, because of the redundancy of the two downgrading vectors, we can simplify 2.17 eliminating one of them. Eliminating W^{+}_t gives:

$$h_1 \int_{x=td_1+\sum_{\tau=1}^{t} W_{1\tau}}^{\infty} \left(x - td_1 - \sum_{\tau=1}^{t} W_{1\tau}\right) f_1(x)dx$$

$$+ \sum_{i=2}^{n-1} h_i \int_{x=td_i+\sum_{\tau=1}^{t} W_{i\tau}-\sum_{\tau=1}^{t} W_{i-1,\tau}}^{\infty} \left(x - td_i - \sum_{\tau=1}^{t} W_{i\tau}+\sum_{\tau=1}^{t} W_{i-1,\tau}\right) f_i(x)dx \qquad 2.18$$

$$+ h_n \cdot \int_{x=td_n-\sum_{\tau=1}^{t} W_{n-1,\tau}}^{\infty} \left(x - td_n + \sum_{\tau=1}^{t} W_{n-1,\tau}\right) f_n(x)dx$$

where $f_i(x)$ is the p.d.f. for $\sum_{\tau=1}^{t} N_\tau q_{i\tau}$.

To show that downgrading does not reduce the expected holding costs, it suffices to show that the partial derivatives of (2.18) with respect to $W^{-}_{1,t},...,W^{-}_{n,t}$ are all non-negative when evaluated at $W^{-}_t = 0$. For each downgrading variable $W_{i,t}$, for i = 1,...,n, the partial derivative of (2.18) is:

$$\frac{\partial}{\partial W^{-}_{i,t}} = h_{i+1} \int_{x=td_{i+1}+W_{i,t}-W_{i+1,t}}^{\infty} f_{i+1}(x)dx - h_i \int_{x=td_i+W_{i-1,t}-W_{i,t}}^{\infty} f_i(x)dx \qquad 2.19$$

Evaluating (2.19) at $W^{-}_{t} = 0$ for i = 1,...,n and substituting $R = d_i/E[q_i] = d_{i+1}/E[q_{i+1}]$ into the limits of integration gives:

$$\frac{\partial}{\partial W_{i,t}} = h_{i+1}\int_{x=E[q_{i+1}]tR}^{\infty} f_{i+1}(x)dx - h_i\int_{x=E[q_i]tR}^{\infty} f_i(x)dx \qquad 2.20$$

To show that (2.20) is non-negative for all i = 1,...,n and t = 1,...,T, we first note that $f_1(x_1),...f_n(x_n)$ share the same distributional form, although they may have different means and standard deviations. This fact follows from our assumption that the probability distributions for $q_1,...,q_n$ all have the same stable distributional form. Stable distributions are completely specified by their means and variances. A linear combination of random variables from a stable distributional form has the same distributional form. (See Allen, Braswell, and Rao (1974) for a discussion of the properties of stable distributions.)

Let Nq_i and $(N\sigma)_i$ be the mean and standard deviation for the pdf $f_i(x_i)$. It can easily be shown that:

$$\overline{Nq_i} = E\left[\sum_{\tau=1}^{t} N_\tau q_{i+1,\tau}\right] = E[q_{i+1}]\sum_{\tau=1}^{t} N_\tau, \text{ and}$$

$$(N\sigma)_i = \text{Std. Dev}\left[\sum_{\tau=1}^{t} N_\tau q_{i+1,\tau}\right] = \sigma_{i+1}\sqrt{\sum_{\tau=1}^{t} N_\tau^2}.$$

Because the service levels are close to 1, and our N vector satisfies the service level constraints, it follows that the lower limits of integration of both terms in 2.20 are less than the expected values of the respective distributions. That is: $E[q_i]\,tR = td_i \le \overline{Nq_i}$, and $E[q_{i+1}]\,tR = td_{i+1} \le \overline{Nq_{i+1}}$.

For both of the terms of 2.20, the lower limit of integration can be expressed in terms of a "standardized variate", or number of standard deviations below the mean. Let a be the standardized variate that is associated with the lower limit of integration in the first term, and let b be the standardized variate that is associated with the lower limit of integration in the second term.

$$a = \frac{\overline{Nq_{i+1}} - E[q_{i+1}]\,tR}{(N\sigma)_{i+1}}, \text{ and } b = \frac{\overline{Nq_i} - E[q_i]\,tR}{(N\sigma)_i}.$$

We can now express 2.20 in terms of a normalized distribution $f(x)$ which has the same functional form as $f_1(x_1),\ldots,f_n(x_n)$ with mean = 0 and standard deviation = 1:

$$\frac{\partial}{\partial W_{i,t}} = h_{i+1}\int_{x=-a}^{\infty} f(x)dx - h_i\int_{x=-b}^{\infty} f(x)dx \qquad 2.21$$

Because $h_{i+1} \geq h_i$, expression 2.21 is positive if $a \geq b$. To see that this is the case, recall that, by assumption, $cv_i \geq cv_{i+1}$, or $(cv_i)^{-1} \leq (cv_{i+1})^{-1}$. We can multiply through by a constant to obtain:

$$\frac{\left(\sum_{\tau=1}^{t} N_\tau - tR\right)}{\sqrt{\sum_{\tau=1}^{t} N_\tau^2}}(cv_{i+1})^{-1} \geq \frac{\left(\sum_{\tau=1}^{t} N_\tau - tR\right)}{\sqrt{\sum_{\tau=1}^{t} N_\tau^2}}(cv_i)^{-1}.$$

After substituting for cv_i and cv_{i+1}, we have:

$$\frac{E[q_{i+1}]\left(\sum_{\tau=1}^{t} N_\tau - tR\right)}{(N\sigma)_{i+1}} \geq \frac{E[q_i]\left(\sum_{\tau=1}^{t} N_\tau - tR\right)}{(N\sigma)_i}.$$

Or equivalently:

$$a = \frac{\overline{Nq_{i+1}} - E[q_{i+1}]\,tR}{(N\sigma)_{i+1}} \geq \frac{\overline{Nq_i} - E[q_i]\,tR}{(N\sigma)_i} = b\,.$$

It follows that the gradient of total holding costs with respect to downgrading from i to j = i+1 is positive for all $W^{-}_{i\tau} > 0$, for $\tau = 1,\ldots, t$. Thus, there will be no downgrading from product i to product j = i+1 in an optimal solution to SCA when conditions 2.14, 2.15, and 2.16 hold and the yield distributions of products i = 1,...,n share the same stable distributional form..◊

This result agrees with intuition. Because the ratio of expected yield to quantity demanded is constant for all products, a given value of N provides a higher service level for products with low coefficients of variation. Downgrading from product i to product j = i+1, where $cv_i > cv_j$, would decrease the service level of one product (i) in order to increase that of another product (j) which is already higher. This could only increase total cost because as the service level for a product increases, so does the marginal expected inventory cost.

Claim: Under conditions 2.14, 2.15, and 2.16, there exists an optimal solution to SCAD in which:

$$N_t = \frac{d_1 + K_1\sigma_1\left(\sqrt{\sum_{\tau=1}^{t} N_\tau^2} - \sqrt{\sum_{\tau=1}^{t-1} N_\tau^2}\right)}{E[q_1]}, \quad t = 1,\ldots,T \qquad 2.22$$

where K_1 is the number of standard deviations that are implied by the service level for product 1.

Proof: From the previous result, we know that it is not optimal to downgrade under conditions 2.14, 2.15, and 2.16. When we eliminate the downgrading variables from the deterministic service level constraints 2.11, we have:

$$\text{Prob}\left[\sum_{\tau=1}^{t} N_\tau q_{i\tau} \geq td_i\right] \geq \alpha_i \quad \text{is equivalent to} \quad \sum_{\tau=1}^{t} N_\tau \geq \frac{td_i + K_i\sigma_i\sqrt{\sum_{\tau=1}^{t} N_\tau^2}}{E[q_i]}.$$

Because K_1cv_1 is at least as large as K_icv_i for all products $i = 1,...,n$, any vector $N = \{N_1,...,N_T\}$ that satisfies the above service level constraints for product 1 in all periods $t = 1,...,T$, automatically satisfies the service level constraints for products $i = 2,...,n$. Thus, the solution that is defined in equation 2.22 is feasible in SCAD. It remains to show that this solution is optimal.

Because the objective function penalizes both production and inventory, the constraint for product 1 for the final period T will be tight in any optimal solution to (SCAD):

$$\sum_{\tau=1}^{T} N_\tau = \frac{td_1 + K_1\sigma_1\sqrt{\sum_{\tau=1}^{T} N_\tau^2}}{E[q_1]}$$

Now, let us modify the way in which we express this equality:

$$\sum_{\tau=1}^{T} N_\tau = \frac{td_1 + K\sigma_1\sqrt{\left(\sum_{\tau=1}^{T} N_\tau\right)^2 \sum_{\tau=1}^{T} X_\tau^2}}{E[q_1]} = \frac{td_1 + K_1\sigma_1\left(\sum_{\tau=1}^{T} N_\tau\right)\sqrt{\sum_{\tau=1}^{T} X_\tau^2}}{E[q_1]} \qquad 2.23$$

where $X_t = \frac{N_t}{\sum_{\tau=1}^{T} N_\tau}$, for t = 1,...,T. Some algebraic maniplulations on 2.23 yield:

$$\sum_{\tau=1}^{T} N_\tau = \frac{T\, d_1}{E[q_1] - K_1\sigma_1\sqrt{\sum_{\tau=1}^{T} X_\tau^2}} \qquad 2.24$$

We can now consider the following optimization problem:

$$\text{Min } \mathbf{N} = \frac{T\, d_1}{E[q_1] - K_1\sigma_1\sqrt{\sum_{\tau=1}^{T} X_\tau^2}} = g(X_1, \ldots, X_T) \qquad 2.25$$

$$\text{s.t. } (\mathbf{N},X) \in \Omega, \text{ where } \Omega = \left\{ \mathbf{N}, X \geq 0; \sum_{\tau=1}^{T} X_\tau = 1; \mathbf{N}\sum_{\tau=1}^{t} X_\tau \geq \frac{T\, d_1 + K_1\sigma_1 \mathbf{N}\sqrt{\sum_{\tau=1}^{t} X_\tau^2}}{E[q_1]},\ t = 1,\ldots, T \right\}$$

Optimization problem 2.25 seeks to allocate production to each period in order to minimize the total production (**N**) that is required to satisfy all T of the service level constraints for product 1. **N** represents the sum of the production in periods 1,..., T. X_τ represents the fraction of this production that is allocated to period τ.

Let $X^T = \left[\frac{N_1}{\sum_{\tau=1}^{T} N_\tau}, \ldots, \frac{N_T}{\sum_{\tau=1}^{T} N_\tau}\right]$,

where $N_t = \frac{d_1 + K_1\sigma_1\left(\sqrt{\sum_{\tau=1}^{t} N_\tau^2} - \sqrt{\sum_{\tau=1}^{t-1} N_\tau^2}\right)}{E[q_1]}$, t = 1,...,T.

By construction, the vector X is feasible in Ω. To show that X is optimal, we compute the gradient with respect to g(X):

$$\nabla g(X)^T = \left[\frac{\partial g}{\partial X_1}, \ldots, \frac{\partial g}{\partial X_2}\right] = \frac{T\, d_1}{\left(E[q_1] - K\sigma_1\sqrt{\sum_{\tau=1}^{T} X_\tau^2}\right)^2} \frac{K\sigma_1}{\sqrt{\sum_{\tau=1}^{T} X_\tau^2}} [X_1,\ldots,X_T] \qquad 2.26$$

Let δ be a vector that such that $X + \delta \in \Omega$. Recall that the vector X represents a feasible allocation of total production N among periods 1,...,T such that everything is produced as late as possible without violating the service level constraints. Therefore, any feasible perturbation δ of X must shift production to earlier periods. In other words, there must exist a positive real number ε such that for some time period index k:

$$\sum_{\tau=1}^{k} \delta_\tau = \varepsilon > 0, \text{ and } \sum_{\tau=k+1}^{T} \delta_\tau = -\varepsilon < 0 \qquad 2.27$$

We can now consider the first order optimality condition for the minimization of a convex function over a convex region:

$$\nabla g(X)^T \cdot \delta = \frac{T\, d_1}{\left(E[q_1] - K_1\sigma_1\sqrt{\sum_{\tau=1}^{T} X_\tau^2}\right)^2} \left(\sum_{\tau=1}^{k} X_\tau\delta_\tau + \sum_{\tau=k+1}^{T} X_\tau\delta_\tau\right)$$

$$\geq \frac{T\, d_1}{\left(E[q_1] - K_1\sigma_1\sqrt{\sum_{\tau=1}^{T} X_\tau^2}\right)^2} \left(X_k \sum_{\tau=1}^{k} \delta_\tau + X_{k+1} \sum_{\tau=k+1}^{T} \delta_\tau\right)$$

$$= \frac{T\, d_1}{\left(E[q_1] - K_1\sigma_1\sqrt{\sum_{\tau=1}^{T} X_\tau^2}\right)^2} (X_k - X_{k+1})\,\varepsilon > 0 \qquad 2.28$$

Equation 2.28 follows from the relationship shown in 2.27 between positive real number ε and time period index k and the fact that vector X has been constructed such that $X_i > X_j$ for $i < j$. Thus X is an optimal solution to 2.25, and, under conditions 2.14, 2.15, and 2.16,there exists an optimal solution to SCAD in which:

$$N_t = \frac{d_1 + K\sigma_1\left(\sqrt{\sum_{\tau=1}^{t} N_\tau^2} - \sqrt{\sum_{\tau=1}^{t-1} N_\tau^2}\right)}{E[q_1]}, \quad t = 1,\ldots,T \quad \Diamond$$

Corollary: When conditions 2.14, 2.15, and 2.16 hold, an optimal solution to SCAD will have $N_t \dashrightarrow d_1/E[q_1]$ as $t \dashrightarrow \infty$.

The intuition of the results above can be explained as follows: In each period, the expected production yield exceeds demand by a certain safety margin. Recall from the equations defined in 2.11 that this safety margin is at least as large as K standard deviations of the cummulative production yield, where the standard deviation can be expressed as:

$$\sigma_i\sqrt{\sum_{\tau=1}^{t} N_\tau^2}$$

As t, the length of the planning horizon, increases to infinity, the standard deviation of the cummulative production yield increases, but at a slower and slower rate. Thus, the expected cummulative production approaches cummulative demand in the limit as the length of the planning horizon goes to infinity.

It can also be noted that, for a given amount of total production $N = N_1 + ...+ N_t$, the standard deviation of the cummulative production yield of product i is minimized when production is distributed as evenly as possible over the periods 1,...,t. Thus, it may be desireable to avoid lotsizing decision rules which result in large variances in production quantities from period to period.

Chapter 3: Heuristics

Because the chip co-production problem is difficult to solve optimally in the presence of random yields, it is necessary to resort to heuristics. The heuristics that we describe in this section have been motivated by three things: industry practice, the analytic results described above, and the linear approximations of the service level version of the problem that has been studied by Bitran and Dasu (1988) and Bitran and Leong(1990).

Recall from the original dynamic programming formulation that the chip co-production problem involves two stages in each period. In the first stage, the Morning problem, a lot size is determined. In the second stage, the Afternoon problem, the random yields have been observed, and downdgrading decisions are made. Thus two types of heuristics are necessary, lot sizing heuristics, and downgrading heuristics.

We have studied two types of lot sizing heuristics: single period, myopic rules and multi period, look ahead rules. The first rule that we consider is defined as follows:

L1:

Step 1: Determine a service level $\alpha_i = p_i /(h_i+p_i)$.

Step 2: Compute $\phi_i(1)$: Prob $(q_{it} \geq \phi_i(1)) = \alpha_i$, for all i.

Step 3: Compute NetD(i,t) = Demand(i,t) - Inventory(i,t-1), for all i.

Step 4: $N_t = \underset{i}{MAX} \left\{ \frac{NetD(i,t)}{\phi_i(1)} \right\}$

Note that steps 1 and 2 need to be done only once. In each period it is necessary only to make the calculation in steps 3 and 4. This lot size rule considers only the direct yields of the individual products, it does not account for the effects of downgrading. The following rule attempts to do this by using the aggregate net demand and yield for each product. For example, the aggregate yield of product j is equal to the sum of the yields of all of the products 1,...,j that can be downgraded to product j. Naturally, this aggregate yield must be compared to the aggregate net demand of product j, i.e. the sum of the net demands for products 1,..j. We define the rule as follows:

L2:

Step 1: Determine the service levels $\alpha_i = p_i /(h_i+p_i)$, for all i.

Step 2: Compute $A\phi_i(1)$: Prob ($q_{1t} + ... + q_{it} \geq A\phi_i(1)) = \alpha_i$, for all i.

Step 3: Compute NetD(i,t) = Demand(i,t) - Inventory(i,t-1), for all i.

Step 4: $$N_t = \max_i \left\lceil \frac{\sum_{j=1}^{i} \text{NetD}(j,t)}{A\phi_i(1)} \right\rceil$$

Again it is necessary to perform steps 1 and 2 only once.

Note that in both of these first two lot-sizing rules, the decision is based only upon the net demand for only the next period. This can lead to a large variance in batch sizes from one period to the next, increasing the total variance of production over the planning horizon. For example, suppose that a total of $N = N_1 + N_2$ units are put into the process in two consecutive units. Recall that the yield vectors $\mathbf{q}_1$ and $\mathbf{q}_2$ are composed of random variables. The variance of $N_1\mathbf{q}_1 + N_2\mathbf{q}_2$ is minimized when $N_1 = N_2$

= N/2, and is maximized when either $N_1 = N$, $N_2 = 0$ or $N_1 = 0$, $N_2 = N$. It is therefore desirable to try to smooth production.

The following lotsizing rule attempts to smooth production by considering two periods of demand.

L3:

Step 1: Determine the service levels $\alpha_i = p_i / (h_i + p_i)$, for all i.

Step 2: Compute $A\phi_i(1)$: Prob ($q_{1t} + \ldots + q_{it} \geq A\phi_i(1)$) = α_i, for all i.

Step 3: Compute NetD(i,t) = Demand(i,t) - Inventory(i,t-1), for all i.

Step 4: $$N_t = \underset{i}{MAX} \left\lceil \frac{\sum_{j=1}^{i} (NetD(j,t) + d_{j,t+1})}{2 \cdot A\phi_i(1)} \right\rceil$$

Note that steps 1 and 2 need to be done only once. In each period it is necessary only to make the calculation in steps 3 and 4. As in L2, L3 attempts to capture the effects of downgrading by using aggregate demands and yields for all products downgradeable to i.

L3 differs from L2 in that it is concerned with the service level at the end of the second period of future production instead of at the end of the first. It is conservative in that it uses the single period fractile for the yields. This is equivalent to assuming that the same quantity will be produced in each of the next two periods, and that the yields will be the same in both production batches. However, this ignores the following fact:

> If x_1 and x_2 are two independent identically distributed occurances of the random variable X, and $y \leq E[X]$, then:
>
> $$\text{Prob} (x_1 + x_2 \geq 2y) \geq 2 \text{ Prob} (x_1 \geq y)$$

(Note that this result is widely known as the Central Limit Theorem.)

The following lotsizing rule attempts to smooth production by considering two periods of demand.

L4:

Step 1: Determine the service levels $\alpha_i = p_i/(h_i+p_i)$, for all i.

Step 2: Compute $A\phi_i(2)$: $\text{Prob}\left(\sum_{j=1}^{i}(q_{jt} + q_{j,t+1}) \geq A\phi_i(2)\right) = \alpha_i$, for all i.

Step 3: Compute NetD(i,t) = Demand(i,t) - Inventory(i,t-1), for all i.

Step 4: $N_t = \underset{i}{\text{MAX}}\left\{\dfrac{\sum_{j=1}^{i}(\text{NetD}(j,t) + d_{j,t+1})}{A\phi_i(2)}\right\}$

The next lotsizing rule goes one step further, and considers three periods of demand:

L6:

Step 1: Determine the service levels $\alpha_i = p_i/(h_i+p_i)$, for all i.

Step 2: Compute $A\phi_i(3)$: $\text{Prob}\left(\sum_{j=1}^{i}(q_{jt} + q_{j,t+1} + q_{j,t+2}) \geq A\phi_i(3)\right) = \alpha_i$, for all i.

Step 3: Compute NetD(i,t) = Demand(i,t) - Inventory(i,t-1), for all i.

Step 4: $N_t = \underset{i}{\text{MAX}}\left\{\dfrac{\sum_{j=1}^{i}(\text{NetD}(j,t) + d_{j,t+1} + d_{j,t+2})}{A\phi_i(3)}\right\}$

The heuristics described above, L1, L2, L3, L4 and L6, are for use in solving the "morning" problem. That is, they all determine the size of the batch which should be put into the process. We have also developed another set of heuristics for solving the "afternoon" problem. These are employed after the yields from the process have been realized to determine how much to downgrade when the yields of one or more products are insufficient to satisfy demand. The first rule that we propose is extremely myopic:

D1: For each of the items $i = 2,...,n$.

$J_{i,t} = J_{i,t-1} + N_t q_{it} - d_{it}$.

If $J_{it} \geq 0$ then next i.

Else for $j = i-1$ to 1:
While $(J_{it} < 0)$
$W_{jit} = MAX(0, MIN(J_{jt}, - J_{it}))$
$J_{jt} = J_{jt} - W_{jit}$
$J_{it} = J_{it} + W_{jit}$

where W_{jit} is used to represent the amount downgraded from j to i in period t. Note that this is a slight abuse of our earlier notation where the downgrading variables were defined only for adjacent items.

The effect of L1 is that if there is an insufficient supply of item i, we downgrade from excess inventory of product i-1 until either the shortage of i or the inventory of i-1 is exhausted. If the shortage of i has not been eliminated, we attempt to downgrade from inventories of products i-2, i-3,..., 1, stopping as soon as either the shortage of i or the inventories of all of i's predessesors have been exhausted. This rule never allows an item to go

into backorder unless there is no inventory of any higher level item available. Its shortcoming is that it does not consider the impact that the downgrading decision could have upon the production costs in the following period.

The immediate benefit of downgrading a unit of product i to fill a backorder for product i+1 is $p_{i+1} + h_i$, the sum of the backorder and holding costs that are saved. The cost of downgrading the item is a function of its effect upon the lotsizing problem in the following period. For example, suppose that we are using lotsize rule L1. Downgrading a single unit of item j in period t has the effect of increasing its net demand in period t+1 by one unit. If item j is the one which satisfies the maximization in Step 3 of L1, then the marginal cost of downgrading is $rc/\phi_j(1)$. It follows that if $rc/\phi_j(1) > p_{j+1} + h_j$, it may be not be wise to downgrade indiscriminantly.

The following rule attempts to trade off the increased production costs in the following period against the decrease in the immediate backorder and holding costs before automatically downgrading:

D2: For each of the items i = 2,...,n.

$J_{i,t} = J_{i,t-1} + N_t q_{it} - d_{it}$.

If $J_{it} \geq 0$ then next i.

Else for j = i-1 to 1:

While ($J_{it} < 0$)

IF ($rc/\phi_j(1) < p_i + h_j$) THEN: $W_{jit} = MAX(0, MIN(J_{jt}, -J_{it}))$

ELSE:

$$N_{t+1} = \underset{i}{MAX} \left\lceil \frac{(d_{i,t+1} - J_{it})}{\phi_i(1)} \right\rceil$$

$$W_{jit} = MAX\{0, MIN\{J_{jt}, -J_{it}, N_{t+1}\phi_j(1) - (d_{j,t+1} - J_{jt})\}\}$$

$J_{jt} = J_{jt} - W_{ijt}$
$J_{it} = J_{it} + W_{ijt}$

Note that variations of this rule can be used in conjunction with the other lotsizing rules. The following variation is appropriate when D2 is used with L2:

D2 (to be used with L2): For each of the items i = 2,...,n.

$J_{i,t} = J_{i,t-1} + N_t q_{it} - d_{it}$.

If $J_{it} \geq 0$ then next i.

Else for j = i-1 to 1:

While ($J_{it} < 0$)

IF ($rc/A\phi_j(1) < p_i + h_j$) THEN: $W_{jit} = MAX(0, MIN(J_{jt}, -J_{it}))$

ELSE:

$$N_{t+1} = \underset{i}{MAX} \left\{ \frac{\sum_{j=1}^{i} (d_{j,t+1} - J_{jt})}{A\phi_i(1)} \right\}$$

$$W_{jit} = MAX\left\{0, MIN\left\{J_{jt}, -J_{it}, N_{t+1}A\phi_j(1) - \sum_{k=1}^{j} (d_{k,t+1} - J_{kt})\right\}\right\}$$

$J_{jt} = J_{jt} - W_{ijt}$
$J_{it} = J_{it} + W_{ijt}$

The heuristics that are described in this section are intended to be used for solving the lotsizing and downgrading problems as they arise in practice. In the following section, we discuss a lower bound on the expected cost of an optimal solution to the original dynamic program. By simulating the performance of the heuristics in a random yield environment, it is possible to statistically estimate the expected costs of using them. We can

evaluate the performance of the heuristics relative to the lower bound. In Section 5, we discuss specific Monte Carlo simulations that have been used to test the heuristics.

Chapter 4: A Lower Bound for the Dynamic Program

In order to evaluate the performance of heuristics, it is necessary to obtain a lower bound on the value of an optimal solution to the original dynamic programming formulation of the co-production planning problem. The bound that we present belongs to one of the classes of bounds that are discussed by Birge (1982).

Recall the original dynamic program, problem (S), which was presented in Chapter 1. In this formulation, $M_t(\mathbf{J}_{t-1})$ represents the lot sizing decision that will made at the beginning (in the "morning") of period t. $A_t(\mathbf{J}_{t-1},N_t,\mathbf{q}_t)$ represents the downgrading decisions that will be made after the yield vector $\mathbf{q}_t$ is observed in period t (in the "afternoon"). The problem is difficult to solve because of the randomness that is associated with these yield vectors. Lotsizing and downgrading decisions must consider not only the direct effects of the random yields, but also the secondary effects of future decisions which will be made after some of the random yield outcomes are observed.

Recall from our discussion of the original dynamic program that, in practice, the structure of the costs generally discourages downgrading that would increase the backorders of a product in the current period. That is, the backorder costs are generally non-decreasing in the stringency of product specifications. Thus, the original formulation can usually be simplified by ignoring constraints 1.6.

Claim: Consider the modified version of the dynamic program $M_1(\mathbf{J}_0)$ in which constraints 1.6 are ignored. If the backorder and holding costs are

non-increasing and non-decreasing in the product indices, , then there exists an optimal solution to the modified problem in which:

$$\mathbf{W}_t^- \le \mathrm{Max}(0, \mathbf{J}_{t-1} + \mathbf{q}_t N_t + \mathbf{W}_t^+), \quad \text{for } t = 1,...,T \qquad 1.6$$

Proof: The proof is by contradiction. Suppose that an optimal solution has:

$$\mathbf{W}_{it}^- = \delta - \mathrm{Max}(0, \mathbf{J}_{i,t-1} + \mathbf{q}_{it} N_t + \mathbf{W}_{it}^+), \quad \text{for some } i,t, \text{ and } \delta > 0. \qquad 4.1$$

Let Z be the value of this optimal solution. By equation 4.1, $J_{i,t} < 0$. Without loss of generality, let us assume that, the result of downgrading product i was to decrease the backorders of product j (> i) by δ. If we decrease W^-_{it} by δ, the change in the costs would be:

$$(p_j - p_i)\delta + M_{t+1}(J_{1,t}, J_{i,t} + \delta, ..., J_{j,t} - \delta, ...) - M_{t+1}(J_{1,t}, J_{i,t}, ..., J_{j,t}, ...) \qquad 4.2$$

Because $p_j \le p_i$, the first term in 4.2 is non-positive. Recall that $M_{t+1}(\mathbf{J}_t)$ is the optimal cost of periods t+1. ..., T given an initial inventory vector of $\mathbf{J}_t$. Because a solution to $M_{t+1}(J_{1,t}, J_{i,t}, ..., J_{j,t}, ...)$ would always be feasible in $M_{t+1}(J_{1,t}, J_{i,t} + \delta, ..., J_{j,t} - \delta, ...)$, the result of the subtraction in equation 4.2 of the third term from the second is also non-positive. Thus, the value of the modified solution is no larger than Z. If it is equal to Z, then it is an alternate optimum. If it is less than Z, then the optimality of the original solution is contradicted. QED

Claim: Consider the modified version of the dynamic program $M_1(J_0)$ in which constraints 1.6 are ignored. The value of an optimal solution to this problem is at least as low as that for the original.

Proof: The modified version is more tightly constrained than the original. Thus an optimal solution to the original is feasible in the modified version. Because the two problems have the same objective function, the value or an optimal solution to the less constrained problem must be less than or equal to that of the original. QED

Although the modified version is itself very difficult to solve, it facilitates the development of a lower bound. Note that a lower bound on the value of the optimal solution to the modified version is also be a lower bound on the original problem. It is in this spirit that we propose the following lower bound on the cost of an optimal solution to the original stochastic dynamic program $M_1(J_0)$.

$LB1(J_0) =$

$$\int_{q1\text{-}qT} \min_{\substack{N_1,\ldots,N_T \\ W_1^+,W_1^-\ldots,W_T^+,W_T^-}} \left\{ \sum_{t=1}^{T} r^{t-1} \left[cN_t + \sum_{i=1}^{n} h_i J_{it}^{+} + \sum_{i=1}^{n} p_i J_{it}^{-} \right] \right\} f(q_1,..,q_T) dq_1,..,dq_T \qquad 4.1$$

such that:

$$J_0 - J_t^{+} + \sum_{\tau=1}^{t} (N_\tau q_\tau + W_\tau^{+} - W_\tau^{-}) \le \sum_{\tau=1}^{t} d_\tau, \quad \text{for } t = 1,\ldots,T \qquad 4.2a$$

$$-J_0 - J_t^{-} - \sum_{\tau=1}^{t} (N_\tau q_\tau + W_\tau^{+} - W_\tau^{-}) \le \sum_{\tau=1}^{t} -d_\tau, \quad \text{for } t = 1,\ldots,T \qquad 4.2b$$

$$W_{it}^{-} = W_{i+1,t}^{+}, \text{ for } i = 1,\ldots,n-1 \text{ and } t = 1,\ldots,T \qquad 4.3a$$

$W_{nt}^{+} = W_{1t}^{-} = 0$, for t = 1,...,T 4.3b

$W_{it}^{+} \geq 0, W_{it}^{-} \geq 0$, for i = 1,...,n-1 and t = 1,...,T 4.4a

$N_t \geq 0$, for t = 1,...,T 4.4b

Note that the term inside of the integration is a linear program in 3nT variables and 2nT constraints.

Claim: $LB1(\mathbf{J}_0)$ is a lower bound on the value of an optimal solution to problem $M_1(\mathbf{J}_0)$.

Proof: In order to prove the claim, we need only show that $LB1(\mathbf{J}_0)$ is a lower bound on the value of an optimal solution to the version of the original dynamic program in which constraints 1.6 are ignored. This can be shown by induction on the number of periods in the horizon. Consider first a single period problem:

$$M_T(\mathbf{J}_{T-1}) = \underset{N_T}{\text{Min}} \left\{ cN_T + \int_{\mathbf{q}_T} \underset{W_T^{+}, W_T^{-}}{\text{Min}} \left\{ \sum_{i=1}^{n} h_i J_{iT}^{+} + \sum_{i=1}^{n} p_i J_{iT}^{-} \right\} f(\mathbf{q}_T)\, d\mathbf{q}_T \right\} \quad 4.5$$

where:

$J_{iT}^{+} = \text{Max}\left(0, J_{i,T-1} + N_T q_{iT} + W_{iT}^{+} - W_{iT}^{-} - d_{iT}\right)$, for i = 1,...,n. 4.5a

$J_{iT}^{-} = \text{Max}\left(0, -J_{i,T-1} - N_T q_{iT} - W_{iT}^{+} + W_{iT}^{-} + d_{iT}\right)$ for i = 1,...,n. 4.5b

$W_T^{-} \leq \text{Max}\left(0, \mathbf{J}_{T-1} + \mathbf{q}_T N_T + W_T^{+} - \mathbf{d}_T\right)$ 4.5c

$N_T \geq 0$ 4.5d

$W_{iT}^{-} = W_{i+1,T}^{+}$, for i = 1,...,n-1 4.5e

$W_{iT}^{+} \geq 0, W_{iT}^{-} \geq 0$, for i = 1,...,n-1 4.5f

$W_{nT}^{+} = W_{1T}^{-} = 0$ 4.5g

Note that we can move the first term in 4.5 inside of the integral to obtain the following equivalent expression:

$$M_T(J_{T-1}) = \min_{N_T}\left\{\int_{q_T} cN_T + \min_{W_T^+, W_T^-}\left\{\sum_{i=1}^{n} h_i J_{iT}^+ + \sum_{i=1}^{n} p_i J_{iT}^-\right\} f(q_T)\, dq_T\right\}$$

$$= \min_{N_T}\left\{\int_{q_T} \min_{W_T^+, W_T^-}\left\{cN_T + \sum_{i=1}^{n} h_i J_{iT}^+ + \sum_{i=1}^{n} p_i J_{iT}^-\right\} f(q_T)\, dq_T\right\} \qquad 4.6$$

Let N^* be an optimal solution to problem 4.6. By the definition of optimality, we have, for any random vector q_T :

$$\min_{W_T^+, W_T^-}\left\{cN^* + \sum_{i=1}^{n} h_i J_{iT}^+ + \sum_{i=1}^{n} p_i J_{iT}^-\right\}$$

$$\geq \min_{N_T, W_T^+, W_T^-}\left\{cN_T + \sum_{i=1}^{n} h_i J_{iT}^+ + \sum_{i=1}^{n} p_i J_{iT}^-\right\} \qquad 4.7$$

Where J_{iT}^+ and J_{iT}^- are functions of q_T, N_T, W_{iT}^+, and W_{iT}^- as described above. By integrating both sides of 4.7 with respect to q_T :

$$\int_{q_T} \min_{W_T^+, W_T^-}\left\{cN^* + \sum_{i=1}^{n} h_i J_{iT}^+ + \sum_{i=1}^{n} p_i J_{iT}^-\right\} f(q_T) dq_T \geq$$

$$\int_{q_T} \min_{N_T, W_T^+, W_T^-}\left\{cN_T + \sum_{i=1}^{n} h_i J_{iT}^+ + \sum_{i=1}^{n} p_i J_{iT}^-\right\} f(q_T) dq_T \qquad 4.8$$

The left hand side of 4.8 is an optimal solution to the single period dynamic program. The right hand side is LB1. Thus, LB1 is a lower bound on the value of an optimal solution to a single period problem.

It remains to be shown that if LB1 is a valid lower bound on the cost of a t period problem, then it is also valid for a t+1 period problem. Our inductive assumption is as follows: Given an initial inventory vector of $\mathbf{J}_{T-t}$, $LB1(\mathbf{J}_{T-t})$ is a lower bound on $M_{T-t+1}(\mathbf{J}_{T-t})$, the value of an optimal solution to the t period problem. Using this assumption, we have:

$$M_{T-t}(\mathbf{J}_{T-t-1}) = \min_{N_{T-t}} \left\{ cN_{T-t} + \int_{q_{T-t}} [A_{T-t}(\mathbf{J}_{T-t-1}, N_{T-t}, q_{T-t})] f(q_{T-t})\, dq_{T-t} \right\}$$

$$= \min_{N_{T-t}} \left\{ c N_{T-t} + \int_{q_{T-t}} \left[\min_{W^+_{T-t}, W^-_{T-t}} \left\{ \sum_{i=1}^{n} h_i J^+_{i,T-t} + \sum_{i=1}^{n} p_i J^-_{i,T-t} + r M_{T-t+1}(\mathbf{J}_{T-t}) \right\} \right] f(q_{T-t}) dq_{T-t} \right\}$$

$$\geq \min_{N_{T-t}} \left\{ cN_{T-t} + \int_{q_{T-t}} \left[\min_{W^+_{T-t}, W^-_{T-t}} \left\{ \sum_{i=1}^{n} h_i J^+_{i,T-t} + \sum_{i=1}^{n} p_i J^-_{i,T-t} + r LB1(\mathbf{J}_{T-t}) \right\} \right] f(q_{T-t}) dq_{T-t} \right\} \qquad 4.9$$

As in the single period case, we can shift the first term inside of the integral to obtain:

$$\underset{N_{T-t}}{\text{Min}}\left\{\int_{q_{T-t}}\left[cN_{T-t}+\underset{W^{+}_{T-t},W^{-}_{T-t}}{\text{Min}}\left\{\sum_{i=1}^{n}h_iJ^{+}_{i,T-t}+\sum_{i=1}^{n}p_iJ^{-}_{i,T-t}+rLB1(J_{T-t})\right\}\right]f(q_{T-t})dq_{T-t}\right\}$$

$$=\underset{N_{T-t}}{\text{Min}}\left\{\int_{q_{T-t}}\underset{W^{+}_{T-t},W^{-}_{T-t}}{\text{Min}}\left\{cN_{T-t}+\sum_{i=1}^{n}h_iJ^{+}_{i,T-t}+\sum_{i=1}^{n}p_iJ^{-}_{i,T-t}+rLB1(J_{T-t})\right\}f(q_{T-t})dq_{T-t}\right\} \qquad 4.10$$

Let N^* be an optimal solution to 4.10. For any random vector q_{T-t}, we have, by the definition of optimality.

$$\underset{W^{+}_{T-t},W^{-}_{T-t}}{\text{Min}}\left\{cN^{*}+\sum_{i=1}^{n}h_iJ^{+}_{i,T-t}+\sum_{i=1}^{n}p_iJ^{-}_{i,T-t}+rLB1(J_{T-t})\right\}$$

$$\geq\underset{N_{T-t},\,W^{+}_{T-t},W^{-}_{T-t}}{\text{Min}}\left\{cN_{T-t}+\sum_{i=1}^{n}h_iJ^{+}_{i,T-t}+\sum_{i=1}^{n}p_iJ^{-}_{i,T-t}+rLB1(J_{T-t})\right\} \qquad 4.11$$

Integrating both sides of 4.11 with respect to q_{T-t} gives:

$$\underset{N_{T-t}}{\text{Min}}\left\{\int_{q_{T-t}}\underset{W^{+}_{T-t},W^{-}_{T-t}}{\text{Min}}\left\{cN_{T-t}+\sum_{i=1}^{n}h_iJ^{+}_{i,T-t}+\sum_{i=1}^{n}p_iJ^{-}_{i,T-t}+rLB1(J_{T-t})\right\}f(q_{T-t})dq_{T-t}\right\}$$

$$=\int_{q_{T-t}}\underset{W^{+}_{T-t},W^{-}_{T-t}}{\text{Min}}\left\{cN^{*}+\sum_{i=1}^{n}h_iJ^{+}_{i,T-t}+\sum_{i=1}^{n}p_iJ^{-}_{i,T-t}+rLB1(J_{T-t})\right\}f(q_{T-t})dq_{T-t}$$

$$\geq\int_{q_{T-t}}\underset{N_{T-t},\,W^{+}_{T-t},W^{-}_{T-t}}{\text{Min}}\left\{cN_{T-t}+\sum_{i=1}^{n}h_iJ^{+}_{i,T-t}+\sum_{i=1}^{n}p_iJ^{-}_{i,T-t}+rLB1(J_{T-t})\right\}f(q_{T-t})dq_{T-t}$$

$$=LB1(J_{T-t-1}). \qquad 4.12$$

Obtaining this lower bound is equivalent to evaluating the expectation (with respect to the yield vectors $\mathbf{q}_1,...,\mathbf{q}_T$) of the optimal cost of the following linear program:

$$\text{Min} \sum_{t=1}^{T} r^{t-1}\left[cN_t + \sum_{i=1}^{n} h_i J_{it}^{+} + \sum_{i=1}^{n} p_i J_{it}^{-} \right] \qquad 4.13$$

such that:

$$\mathbf{J}_0 - \mathbf{J}_t^{+} + \sum_{\tau=1}^{t} \left(N_\tau \mathbf{q}_\tau + \mathbf{W}_\tau^{+} - \mathbf{W}_\tau^{-}\right) \le \sum_{\tau=1}^{t} \mathbf{d}_\tau, \quad \text{for } t = 1,...,T \qquad 4.13a$$

$$-\mathbf{J}_0 - \mathbf{J}_t^{-} - \sum_{\tau=1}^{t} \left(N_\tau \mathbf{q}_\tau + \mathbf{W}_\tau^{+} - \mathbf{W}_\tau^{-}\right) \le -\sum_{\tau=1}^{t} \mathbf{d}_\tau, \quad \text{for } t = 1,...,T \qquad 4.13b$$

$$W_{it}^{-} = W_{i+1,t}^{+}, \text{ for } i = 1,...,n-1 \text{ and } t = 1,...,T \qquad 4.13c$$

$$W_{nt}^{+} = W_{1t}^{-} = 0, \text{ for } t = 1,...,T \qquad 4.13d$$

$$W_{it}^{+} \ge 0, W_{it}^{-} \ge 0, \text{ for } i = 1,...,n-1 \text{ and } t = 1,...,T \qquad 4.13e$$

$$N_t \ge 0, \text{ for } t = 1,...,T \qquad 4.13f$$

This linear program has 3nT variables and 2nT constraints. Because the distributions for the yields are continuous, an exact computation of the expected optimal cost of 4.13 is quite difficult. However we can obtain a statistical esimate of LB1 by using Monte Carlo simulation. This estimation procedure involves iteratively generating sets of the vectors $\mathbf{q}_1,...,\mathbf{q}_T$ from the appropriate distributions, and solving 4.29 for each set. Each time a set of yield vectors is generated, and the linear program is solved represents a single sample.. By the Central Limit Theorem, as the number of these samples goes to infinity, our estimate converges in

expectation to LB1. We can use statistical estimation techniques to determine the appropriate number of samples for any desired accuracy.

The following section describes the Monte Carlo simulations that are used to evaluate the performance of the heuristics that are developed in Chapter 3 against the above lower bound.

Chapter 5: Computational Results

In order to evaluate the performance of the heuristics that are described in Chapter 2, we used Monte Carlo simulation. We simulated the performance of each of the pairing of lotsizing and downgrading heuristics in an environment with random yields. By measuring the costs of production, backorders, and inventory over many simulations, we were able to obtain a statistical estimate of the expected total cost of each of the heuristic pairs.

We also used Monte Carlo simulation to determine a statistical estimate of our lower bound. Recall that the lower bound is the expected optimal value of a linear program in which the coefficients in the constraint matrix are random. We "estimated" this expected optimal value by repeatedly generating realizations of the random coefficients and solving the resulting linear programs. This "estimate" of the lower bound on the expected cost of an optimal solution to the original dynamic program provides a benchmark against which we can evaluate the performance of the various heuristic pairs.

In the Monte Carlo simulations, we assumed that the distributions of the conditional-aggregate yields of the items can be represented by the Beta distributions. The conditional -aggregate yield of a given product i is the fraction of items that satisfy the specification for item i+1 which also satisfy the specification for item i. Suppose for example, that there are three serially nested products. The fraction of a production batch which satisfies the loosest tolerences, i.e. grade 3, is a Beta distributed random variable, say β_3. The fraction of of that fraction which also satisfies the next tighter

specification is another Beta distributed random variable, say β_2. Still another Beta distributed random variable, β_1, describes the fraction of those items inside of item 2's specifications which also satisfy the specifications for item 1. If β_1, β_2, and β_3 are three independent random variables representing conditional-aggregate yields, the actual yields of items 1, 2, and 3 would be:

$$q_1 = \beta_1 \beta_2 \beta_3$$

$$q_2 = \beta_2 \beta_3 - q_1$$

$$q_3 = \beta_3 - q_1 - q_2$$

Beta distributions have been widely used in the literature to represent uncertain yields. By setting the parameters appropriately, the first and second moments can be set to match almost any situation we have observed in practice. By using the Beta distributions to represent conditional-aggregate yields as described above, we are also able to capture the correlations between yields of various products that are common in practice.

We ran the simulations with several different sets of data. One of the objectives of the investigation was to determine the robustness of the heuristics with respect to:

- Relative costs of production, holding inventory, and backordering.
- Ratios of demand to expected yields for different products.
- Coefficients of variation in yields.

We attempted to determine the conditions under which various combinations of lotsizing and downgrading rules perform well.

For the first series of tests, we assumed that the cost structure is the following:

- Production Cost: \$8 / unit for products 1, 2, and 3.

- Holding Cost: $1 / unit for products 1, 2, and 3.
- Backorder Cost: $19 / unit for products 1, 2, and 3.

These costs are representative of the situation in the facility that motivated our research. Note the absense of a discount factor. Because the time periods that we observed are generally on the order of a day or a week, the discount factor does not play a significant role in the total costs. For this reason, we implicitly assumed that it is zero.

We considered six different sets of yield distributions, the parameters of which are given in Tables 5.1 and 5.2. In each of these tables, three related sets of yield distributions are defined. The first two columns show three different settings of the α_1 and α_2 for each of the three beta distributed conditional-aggregate yields described above. The next two columns indicate the expected values of β_i and q_i . Finally, the last two columns provide indications as to the level of uncertainty associated

Yield Dist.	i	α_1	α_2	$E[\beta_i]$	$E[q_i]$	90% CI	cv_i
1A	1	3	7	0.3	0.162	(42,317)	0.546
HIGH cv	2	6	4	0.6	0.378	(175,610)	0.336
	3	9	1	0.9	0.36	(136,593)	0.384
1B	1	30	70	0.3	0.162	(118,211)	0.176
MED cv	2	60	40	0.6	0.378	(312,448)	0.110
	3	90	10	0.9	0.36	(288,434)	0.126
1C	1	300	700	0.3	0.162	(148,177)	0.060
LOW cv	2	600	400	0.6	0.378	(357,399)	0.030
	3	900	100	0.9	0.36	(337,384)	0.040

Table 5.1

with the yields. The column titled "90% CI" represents a 90% confidence interval on the number of units in a batch of 1000 that meet the specification

for product i, but not for product i-1. For example, in Yield distribution 1A, we would be 90% confident that between 42 and 317 of the units in a batch of 1000 would meet the specification for product 1. Similarly, we would be 90% confident that between 175 and 610 units would meet the specification for product 2, but not that for product 1. The column titled "cv_i" contains the coefficient of variation for the yields q_i. Note that the expected values of the yield vectors are the same across Yield Distributions 1A, 1B, and 1C, but the coefficients of variation decrease. Similarly, the expected values of the yield vectors are the same across Distributions 2A, 2B, and 2C, but the coeficients of variation are decreasing.

Yield Dist.	i	α_1	α_2	$E[\beta_i]$	$E[q_i]$	90% CI	cv_i
2A	1	3	3	0.5	0.25	(73,502)	0.494
HIGH cv	2	6	3	0.67	0.25	(69,476)	0.494
	3	6	2	0.75	0.25	(71,477)	0.494
2B	1	30	30	0.5	0.25	(184,316)	0.162
MED cv	2	60	30	0.67	0.25	(186,317)	0.162
	3	60	20	0.75	0.25	(184,313)	0.162
2C	1	300	300	0.5	0.25	(229,271)	0.0514
LOW cv	2	600	300	0.67	0.25	(229,272)	0.0514
	3	600	200	0.75	0.25	(229,271)	0.0514

Table 5.2

Notice that, for distributions 1A, 1B, and 1C, the coefficient of variation is much higher for the yield of product 1 than for those of products 2 or 3. In contrast, for distributions 2A, 2B, and 2C, the coefficients of variation are the same for all three products. In practice, the relationship found in the first set of distributions (1A, 1B,and 1C) seems to be more common.

In the first test, the ratio of demand to expected yield is constant for all products. In particular, we have assumed that:

$$\frac{d_i}{E[q_i]} = 1000, \text{ for } i = 1, 2, \text{ and } 3.$$

Thus, for yield distributions 1A, 1B, and 1C, the demands for the three products are 162, 378, and 360 in each period. Similarly, for distributions 2A, 2B, and 2C, the demands are 250, 250, and 250. We have made this assumption because this is the most difficult case for production managers to deal with. Suppose, for example, that the ratio of demand to expected yield is much higher for the most stringently specified product than for the others. In this case, the problem reduces to one of a single product . The less stringently specified products are essentially bi-products and can be ignored in lot sizing decisions. Because the "high grade" product is "difficult" to obtain, downgrading is neither necessary nor desirable. The other case, where the ratio of demand to expected yield is lower for the most stringently specified product than for the others simply does not occur frequently in practice. Because there is usually a trade-off between the cost of the process and the expected yields of "high grade" products, it is rare for production managers to face situations in which they have abundant quantities of these items relative to demand.

Yield Dist.	L2D1	L3D1	L4D1	L6D1	LB
1A	\$44,889 σ_{est}= \$3124	\$43,220 σ_{est}= \$2200	\$39,739 σ_{est}= \$1749	\$42,164 σ_{est}= \$1876	\$36,991 σ_{est}= \$793
1B	\$34,393 σ_{est}= \$445	\$34,403 σ_{est}= \$392	\$34,531 σ_{est}= \$441	\$34,633 σ_{est}= \$338	\$33,156 σ_{est}= \$222
1C	\$32,521 σ_{est}= \$159	\$32,960 σ_{est}= \$148	\$32,736 σ_{est}= \$102	\$32,983 σ_{est}= \$109	\$32,363 σ_{est}= \$62

Demand for products 1, 2, and 3 equals 162, 378, 360 respectively.

Table 5.3

Yield Dist.	L2D2	L3D2	L4D2	L6D2	LB
1A	\$44,625 σ_{est}= \$2893	\$43,033 σ_{est}= \$2133	\$39,358 σ_{est}= \$1683	\$42,136 σ_{est}= \$1869	\$36,991 σ_{est}= \$793
1B	\$34,216 σ_{est}= \$435	\$34,405 σ_{est}= \$390	\$34,215 σ_{est}= \$437	\$34,633 σ_{est}= \$338	\$33,156 σ_{est}= \$222
1C	\$32,742 σ_{est}= \$154	\$32,958 σ_{est}= \$148	\$32,741 σ_{est}= \$155	\$32,983 σ_{est}= \$108	\$32,363 σ_{est}= \$62

Demand for products 1, 2, and 3 equals 162, 378, 360 respectively.

Table 5.4

The performance of the different lotsizing rules can be compared in Tables 5.3 and 5.4. The first two letters of the column headings correspond to the lotsizing rule, and the second two correspond to the downgrading rule which was used. Note that the downgrading rule was held constant in each table. D1 was used to obtain the results in Table 5.3, and D2 was used for table 5.4.

Each entry in these tables contains two values. The upper one is a statisistical estimate, and the lower one is the standard deviation of this estimate. The first four columns of the table contain the estimates (along with their standard deviations) of the expected costs of using the heuristics for four periods while the system is in steady state. The fifth column contains an estimate (and its standard deviation) of the lower bound on the expected cost of an optimal solution to the four period problem.

We obtained these estimates by using Monte Carlo techniques to simulate the performance of the heuristics in the presence of random yields. For each heuristic pair, we repeatedly simulated its operations for eight periods, starting with zero initial inventory. In order to capture the

system in steady state, we measured the costs only over periods 5, 6, 7, and 8. Although the simulations were truncated after period 8, the heuristics were operated as though demand continued indefinitely. Thus, there was no boundary effect in period 8. The values in the tables are based on 40 iterations of this 8 period simulation.

In both Tables 5.3 and 5.4, we can observe that all of the lotsizing rules perform very well vis-a-vis the lowerbound for yield distributions 1B and 1C. For distribution 1B, all of the heuristics result in costs which are within about 4.5% of the lower bound. For 1C, they all result in costs which are within about 1%. Thus, when the coefficients of variation of the yields are low or moderate, all of the heuristics perform well.

However, all the heuristics do not perform uniformly well when the coefficients of variation are high. Although L4D1 and L4D2 get within 7.4% and 6.4% of the lower bound, the other rules do not perform nearly so well. Comparing the estimated expected costs associated with L2D1, L3D1, and L6D1 with that of L4D1, we obtained t-statistics of 1.66, 1.32, and 1.05 respectively. Thus, we can reject the null hypothesis that the costs associated with L4D1 are at least as high as those of L2D1, L3D1, and L6D1 with 95%, 90%, and 85% confidence respectively. The t-statistics for the comparisons of the estimated expected costs associated with L2D2, L3D2, and L6D2 with that of L4D2 are 1.70, 1.39, and 1.36 respectively. We can reject the hypothesis that the costs associated with L4D2 are at least as high as those of L2D2, L3D2, and L6D2 with 96%, 95%, and 91% confidence respectively.

The performance of L4 suggests that there is an advantage associated with considering two periods of demand instead of one. However, the fact that L6 does not perform as well as L4 suggests that it is counter-productive

to consider too many periods of demand. A possible explanation is the following: The consideration of additional period's demand has a smoothing effect upon the lot size. Some smoothing is beneficial in the sense that variability in lot sizes increases to the total variablity of production output. However, too much smoothing "over-damps" the system's ability to respond to extremely low or high yields.

By comparing the results in Table 5.3 to those in 5.4, we observe that, in nearly all cases, the more far-sighted downgrading rule marginally outperforms the myopic one. However, this difference was only significant when lotsizing rule L4 was used. The t-statistic for the comparison of the estimated expected costs of L4D1 to L4D2 was 1.77. Thus, we can reject the null hypothesis that the expected cost of using L4D1 is less than or equal to that of using L4D2 with 96% confidence. The fact that the difference was only significant for lotsizing rule L4 might suggest that the benifit from a far-sighted downgrading rule can only be realized when the lotsizing rule is good.

The results for the second set of distributions, in which the coefficients of variation were the same across the yields of all three products, are given in tables 5.5 and 5.6:

Yield Dist.	L2D1	L3D1	L4D1	L6D1	LB
2A	\$45,913 σ_{est}= \$2676	\$42,036 σ_{est}= \$1762	\$37,630 σ_{est}= \$1051	\$41,880 σ_{est}= \$1786	\$36,062 σ_{est}= \$989
2B	\$34,508 σ_{est}= \$564	\$34,697 σ_{est}= \$352	\$34,286 σ_{est}= \$350	\$34,521 σ_{est}= \$352	\$33,055 σ_{est}= \$219
2C	\$32,673 σ_{est}= \$123	\$32,685 σ_{est}= \$115	\$32,634 σ_{est}= \$106	\$32,966 σ_{est}= \$110	\$32,469 σ_{est}= \$82

Demand for products 1, 2, and 3 equals 250, 250, 250 respectively.

Table 5.5

Yield Dist.	L2D2	L3D2	L4D2	L6D2	LB
2A	\$44,335 σ_{est}= \$2434	\$41,768 σ_{est}= \$1704	\$37,687 σ_{est}= \$1038	\$41,880 σ_{est}= \$1786	\$36062 σ_{est}= \$989
2B	\$34,104 σ_{est}= \$560	\$34,712 σ_{est}= \$437	\$34,286 σ_{est}= \$350	\$34,521 σ_{est}= \$352	\$33,055 σ_{est}= \$219
2C	\$32,673 σ_{est}= \$103	\$32,693 σ_{est}= \$102	\$32,634 σ_{est}= \$106	\$32,966 σ_{est}= \$110	\$32,469 σ_{est}= \$82

Demand for products 1, 2, and 3 equals 250, 250, 250 respectively.

Table 5.6

As in the first set of distributions, L4 was clearly the dominant lot sizing rule. Using it results in expected costs that are within 4.5%, 3.7%, and and 0.5% of the lower bounds for yield distributions 2A, 2B, and 2C respectively. However, there seems to be even less difference between the two downgrading rules here than there is for distributions 1A, 1B, and 1C.

In cases where the ratio of demand to expected yield is lower for the most stringently specified product than for the others, downgrading is "systematic". That is, some downgrading would occur even if the yields were deterministic. Large imbalances of this sort are rare because it is

usually expensive to maintain high yields of the tightly specified products. However, it is not unusual for slight imbalances to be found, particularly in the short term.

To test our heuristics in situations where the ratio of demand to expected yield is lower for the most stringently specified product than for the others, we have modified the original sets of demand that were used with yield distributions 1A, B, and C, and 2A, B, and C. In the first modification, we set:

$$\frac{d_1+50}{E[q_1]} = \frac{d_2}{E[q_2]} = \frac{d_3-50}{E[q_3]} = 1000.$$

Thus, for yield distributions 1A, B, and C, the demands in each period were 112, 378, and 410 respectively. For distributions 2A, B, and C, the demands were 200, 250 and 300. The results of the tests for the first modification are shown in tables 5.7, 5.8, 5.9 and 5.10.

Yield Dist.	L2D1	L3D1	L4D1	L6D1	LB
1A	\$34,260	\$33,537	\$33,942	\$34,355	\$34,303
	σ_{est}= \$1549	σ_{est}= \$1450	σ_{est}= \$731	σ_{est}= \$840	σ_{est}= \$511
1B	\$32,194	\$32,384	\$32,269	\$32,455	\$31,997
	σ_{est}= \$78	σ_{est}= \$56	σ_{est}= \$47	σ_{est}= \$38	σ_{est}= \$88
1C	\$32,148	\$32,202	\$32,113	\$32,133	\$32,044
	σ_{est}= \$22	σ_{est}= \$18	σ_{est}= \$17	σ_{est}= \$114	σ_{est}= \$54

Demand for products 1, 2, and 3 equals 112, 378, 410 respectively.

Table 5.7

Yield Dist.	L2D2	L3D2	L4D2	L6D2	LB
1A	\$34,675 σ_{est}= \$1344	\$33,544 σ_{est}= \$825	\$33,830 σ_{est}= \$643	\$34,346 σ_{est}= \$815	\$34,303 σ_{est}= \$511
1B	\$32,193 σ_{est}= \$78	\$32,384 σ_{est}= \$56	\$32,269 σ_{est}= \$60	\$32,466 σ_{est}= \$68	\$31,997 σ_{est}= \$88
1C	\$32,134 σ_{est}= \$25	\$32,202 σ_{est}= \$18	\$32,113 σ_{est}= \$20	\$32,133 σ_{est}= \$23	\$32,044 σ_{est}= \$54

Demand for products 1, 2, and 3 equals 112, 378, 410 respectively.

Table 5.8

The data in tables 5.7 and 5.8 indicate that the heuristics all perform very similarly when the ratio of demand to expected yield is higher for product 3 than for product 1. Note that all of the heuristics perform very well. Even for the high coefficient of variation (yield distribution 1A), all of the heuristics are all within about 1% of the lower bound. These results support the idea that the problems are easier to solve when the ratio of demand to expected yield is the same across all of the products. We can think of the production system as being driven by the aggregate yield of all of the products that can be used to satisfy the demand for whichever item is in shortest supply. Because the coefficients of variation of the aggregate yields tend to be lower than those for individual yields, the least stable systems are those in which the ratio of demand to expected yield is highest for item 1. Thus these are the systems that are the most difficult , and the most expensive, to manage.

Yield Dist.	L2D1	L3D1	L4D1	L6D1	LB
2A	\$37,398	\$35,733	\$32,672	\$36,639	\$33,614
	σ_{est}= \$1846	σ_{est}= \$1050	σ_{est}= \$596	σ_{est}= \$1029	σ_{est}= \$685
2B	\$32,242	\$32,614	\$32,329	\$32,928	\$32,091
	σ_{est}= \$254	σ_{est}= \$139	σ_{est}= \$110	σ_{est}= \$120	σ_{est}= \$169
2C	\$32,181	\$32,249	\$32,160	\$32,272	\$32,096
	σ_{est}= \$55	σ_{est}= \$36	σ_{est}= \$43	σ_{est}= \$50	σ_{est}= \$55

Demand for products 1, 2, and 3 equals 200, 250, 300 respectively.

Table 5.9

Yield Dist.	L2D2	L3D2	L4D2	L6D2	LB
2A	\$36.707	\$35,667	\$32,650	\$36,639	\$33,614
	σ_{est}= \$1538	σ_{est}= \$1015	σ_{est}= \$575	σ_{est}= \$1029	σ_{est}= \$685
2B	\$32,242	\$32,614	\$32,329	\$32,928	\$32,091
	σ_{est}= \$254	σ_{est}= \$139	σ_{est}= \$107	σ_{est}= \$120	σ_{est}= \$169
2C	\$32,181	\$32,249	\$32,160	\$32,272	\$32,096
	σ_{est}= \$55	σ_{est}= \$36	σ_{est}= \$43	σ_{est}= \$50	σ_{est}= \$55

Demand for products 1, 2, and 3 equals 200, 250, 300 respectively.

Table 5.10

The data in tables 5.9 and 5.10 also support the idea that it is less costly to manage a system in which the item that is in short supply is other than the most stringently specified one. However, in contrast to the data for yield distributions 1A, 1B, and 1C (tables 5.7 and 5.8), here we see that the heuristics do not all perform uniformly well. In fact there seems to be a significant advantage to using lot size rule L4. For the high coeficient of variation case, the estimates of the expected costs of L4D1 and L4D2 are statistically indistinguishable from that of the lower bound. However, the

expected costs of the other heuristics are about 10% to 12% higher than the lower bound.

We also tested a case in which product 2 was the one which was in the shortest supply. In particular, we set:

$$\frac{d_1+50}{E[q_1]} = \frac{d_2-50}{E[q_2]} = \frac{d_3}{E[q_3]} = 1000.$$

Thus, for the first set of yield distributions, the demands in each period were 112, 428, and 360 respectively. For the second set, they were 200, 300 and 250. The results of these tests are given in tables 5.11, 5.12, 5.13, and 5.14.

Yield Dist.	L2D1	L3D1	L4D1	L6D1	LB
1A	\$37,775	\$36,333	\$36,540	\$36,754	\$35,761
	σ_{est}= \$1428	σ_{est}= \$954	σ_{est}= \$802	σ_{est}= \$818	σ_{est}= \$748
1B	\$33,138	\$33,388	\$33,122	\$33,391	\$32,660
	σ_{est}= \$254	σ_{est}= \$199	σ_{est}= \$110	σ_{est}= \$120	σ_{est}= \$157
1C	\$32,441	\$32,537	\$32,436	\$32,552	\$32,265
	σ_{est}= \$76	σ_{est}= \$71	σ_{est}= \$43	σ_{est}= \$56	σ_{est}= \$55

Demand for products 1, 2, and 3 equals 112, 428, 360 respectively.

Table 5.11

Yield Dist.	L2D2	L3D2	L4D2	L6D2	LB
1A	$37,207	$36,458	$36,518	$36,691	$35,761
	σ_{est}= $1281	σ_{est}= $937	σ_{est}= $783	σ_{est}= $812	σ_{est}= $748
1B	$33,138	$33,388	$33,122	$33,391	$32,660
	σ_{est}= $254	σ_{est}= $199	σ_{est}= $110	σ_{est}= $120	σ_{est}= $157
1C	$32,441	$32,537	$32,436	$32,552	$32,265
	σ_{est}= $76	σ_{est}= $71	σ_{est}= $43	σ_{est}= $56	σ_{est}= $55

Demand for products 1, 2, and 3 equals 112, 428, 360 respectively.

Table 5.12

Yield Dist.	L2D1	L3D1	L4D1	L6D1	LB
2A	$37,650	$38,045	$34,918	$38,114	$34,708
	σ_{est}= $1778	σ_{est}= $1242	σ_{est}= $879	σ_{est}= $1029	σ_{est}= $800
2B	$33,019	$33,411	$33,160	$33,636	$32,658
	σ_{est}= $326	σ_{est}= $235	σ_{est}= $195	σ_{est}= $200	σ_{est}= $211
2C	$32,453	$32,550	$32,422	$32,513	$32,328
	σ_{est}= $70	σ_{est}= $71	σ_{est}= $69	σ_{est}= $80	σ_{est}= $70

Demand for products 1, 2, and 3 equals 200, 300, 250 respectively.

Table 5.13

Yield Dist.	L2D2	L3D2	L4D2	L6D2	LB
2A	$38,328	$38,101	$34,918	$38,114	$34,708
	σ_{est}= $1794	σ_{est}= $1200	σ_{est}= $879	σ_{est}= $1029	σ_{est}= $800
2B	$33,019	$33,411	$33,160	$33,636	$32,658
	σ_{est}= $326	σ_{est}= $235	σ_{est}= $195	σ_{est}= $200	σ_{est}= $211
2C	$32,453	$32,550	$32,422	$32,513	$32,328
	σ_{est}= $70	σ_{est}= $71	σ_{est}= $69	σ_{est}= $80	σ_{est}= $70

Demand for products 1, 2, and 3 equals 200, 300, 250 respectively.

Table 5.14

From Tables 5.11, 5.12, 5.13, and 5.14, we can see that L4D1 and L4D2 once again dominate the other heuristics. It is also of interest to compare the costs for this demand scenario with the other two that we have tested. In particular, the costs are highest when the ratio of demand to expected yield is constant across all three products. The costs are somewhat lower when item 2 is the one which is "rare" relative to its demand, and lower yet when item 3 is the "rare" one. This is consistent with the intuition that the lower the rare item is within the product hierarchy, the more flexibility there is. That is, it becomes easier to adjust to shortages by downgrading.

In order to test the heuristics under a different cost structure, we decreased our original backorder cost from \$19 to \$9. Thus the modified cost structure is as follows:

- Production Cost: \$8 / unit for products 1, 2, and 3.
- Holding Cost: \$1 / unit for products 1, 2, and 3.
- Backorder Cost: \$9 / unit for products 1, 2, and 3.

To test the heuristics with this cost structure, we revisited distributions 1A, 1B, and 1C, and assumed that the ratio of demand to expected yields is

Yield Dist.	L2D1	L3D1	L4D1	L6D1	LB
1A	\$43,476	\$42,283	\$39,628	\$41,957	\$34,530
	σ_{est}= \$2807	σ_{est}= \$2026	σ_{est}= \$1744	σ_{est}= \$1775	σ_{est}= \$454
1B	\$34,226	\$34,210	\$34,450	\$34,397	\$32,411
	σ_{est}= \$434	σ_{est}= \$368	σ_{est}= \$426	σ_{est}= \$317	σ_{est}= \$122
1C	\$32,730	\$32,882	\$32,691	\$32,894	\$32,130
	σ_{est}= \$154	σ_{est}= \$143	σ_{est}= \$96	σ_{est}= \$103	σ_{est}= \$33

Demand for products 1, 2, and 3 equals 162, 378, 360 respectively.

Table 5.15

Yield Dist.	L2D2	L3D2	L4D2	L6D2	LB
1A	$43,886	$41,526	$38,915	$41,817	$34,530
	σ_{est}= $2663	σ_{est}= $1864	σ_{est}= $1570	σ_{est}= $1770	σ_{est}= $454
1B	$34,098	$34,213	$34,443	$34,391	$32,411
	σ_{est}= $428	σ_{est}= $366	σ_{est}= $425	σ_{est}= $314	σ_{est}= $122
1C	$32,732	$32,868	$32,688	$32,880	$32,130
	σ_{est}= $151	σ_{est}= $143	σ_{est}= $97	σ_{est}= $100	σ_{est}= $33

Demand for products 1, 2, and 3 equals 162, 378, 360 respectively.

Table 5.16

constant across all three products. The results of these tests are shown in Tables 5.15 and 5.16.

These results are similar to the previous ones. All of the heuristics perform very well vis-a-vis the lower bound when the coeficient of variation of the yields is either moderate or low, distributions 1B and 1C. When the coefficients of variation are high, lotsizing rule L4 dominates the others, and it performs slightly better when used in conjuction with the far-sighted downgrading rule D2.

However, it is also interesting to note that, with the modified cost structure, the best heuristic, L2D2, results in costs that are 12% higher than the lower bound. This is not nearly as good as the 6.4% that was attained with the original cost structure. A possible explanation could be the following: In the original cost structure, backorder costs were about 2.5 times as high as unit production costs. In the modified cost structure, they were only 1.125 times as high. Recall that the lower bound is based on the benefit of omniscience. That is, the lower bound reflects the expected costs of running the system if all decisions could be made with advance

knowledge about the realizations of the random yield outcomes. Because the costs of holding inventory are low relative to those of production and backorders, if the yield realizations are higher in the early periods than in later ones, the omnicient decision maker can produce a lot in these periods and then hold inventory for subsequent periods. In contrast, due to the fact that backorders are generally very expensive, it rarely benifits the omniscient decision maker to satisfy demands in the early periods with later production. Thus, when the yield realizations are higher in the later periods, omniscience provides little advantage. As the costs of backorders are reduced, the omniscient decision maker becomes better able to take advantage of high yield realizations that occur late in the horizon. He can produce a lot in periods with high yields and either backorder or hold inventory for other periods. In other words, the lower the costs of inventory and backorders, the more benefit there is in having advance knowledge of the yield outcomes. As a result, the lower bound is not as tight when these costs are low.

Chapter 6: Discussion

The management of co-production processes in the presence of random yields is an important problem that has received surprisingly little attention in the literature. In many manufacturing environments, a single process produces multiple products simultaneously. Such situations frequently occur when the various products are differentiated from one another by some quantitative measure of performance. For example, the specifications for two semi-conductor chips (A and B) may be identical to one another with the exception that they must operate at different speeds. Because the individual chips in a single process batch may perform at different speeds, some of them may meet the specification for A, while others may meet the specification for B. When there is uncertainty associated with the proportions of the production batch that will fall into the various product categories, managing these co-production processes becomes quite difficult. It becomes even more difficult when there are opportunities to substitute one product for another.

We have formulated the problem as a dynamic program in which the objective is to minimize the expected costs of meeting contractual obligations. In each production period, two decisions are made. After a lotsizing decision is made, the process is run, and the random yields are observed. At this point the output must be either allocated to customers or stored as inventory for future periods. We describe a method for obtaining a lower bound on the optimal solution to this model, and propose several heuristics for solving the problem in practice.

The results of Monte Carlo simulations of our proposed heuristics indicate that the choice of a lot-sizing heuristic has a much more

significant effect upon costs than does the choice of a downgrading rule. We have shown that, for a cost structure that is similar to what we have observed in practice, one of our lotsizing heuristics (L4D2) performs much better than the others, and that it performs very well vis-a-vis the lower bound.

Although these results are encouraging, there are many opportunities for further research. These opportunities can be found by both generalizing the assumptions and broadening the scope of the problem. We have assumed that the yield distributions are parameters of the problem. In practice, management is likely to have some control over these distributions. For example, in semiconductor manufacturing, it may be possible to improve the expectation and variance of the yields by using better materials, decreasing the impurities in the manufacturing environment, or gaining better control over temperatures. An improved understanding of how the costs of running the process are influenced by the yield distributions would aid managers in making process change decisions.

We have also assumed that the yield distributions are well understood by the production manager. In practice, this is often not the case. New technologies replace one another so rapidly that each observation of the current process provides valuable information about the yield distributions. Often, it is desirable to run smaller batches simply to speed up the learning process. This trade-off between the costs of more frequent set-ups and the benefits of more rapid learning needs to be modeled. One method of capturing this trade-off would be to embed a Bayesian update of the yield distributions in each stage of the dynamic

programming formulation that we have proposed. However, this would not be a trivial modification.

Another area for future research is the relationship between production and downgrading decisions and the marketing of the various products. We have assumed that the contractual obligations to supply the products are given a priori, and that the objective is to minimize the expected costs of living up to these obligations. In many situations, demand is not known in advance and maximization of expected profit is a more suitable objective function. A number of interesting research questions arise in this domain. For example, how should pricing decisions be made? What is the best number of product categories to distinguish? Does downgrading have an adverse effect upon demand for the highest grade products? etc.

The problem of managing co-production processes in the presence of random yields is both important and intriguing. Further study of the issues described in this paper can potentially provide significant benefits to a wide variety of both manufacturing and service industries. Because of the wide variety of disciplines that can be brought to bear on the solution of the co-production problem, it should provide a rich frontier for further research.

PART TWO: Managing Hotel Reservations with Uncertain Arrivals
Chapter 7: The Short Term Hotel Problem

The problem that hotel managers face in managing the demand for reservations is quite difficult. They must respond to requests for a variety of types of reservations in order to balance the expected loss of revenue from unsold rooms against both the tangible and intangible costs of "walking" customers, i.e. failing to honor reservations. Since customers who have booked rooms may either cancel or fail to show up with some probability, it is not unusual for hotels to "overbook", i.e. to accept more reservations than they have rooms. Doing this successfully requires a thorough understanding of market dynamics and consumer behavior in several different segments of the market.

This "yield management" problem is similar in concept to the problem of co-production of substitutable products with random yields. In the co-production problem, several different grades of an item are produced simultaneously in a single process. Often, there is uncertainty as to the relative quantities of the various grades in any given batch. Although there are demands for each grade, demand for a particular grade can be satisfied with a higher grade item. Managing such a process involves two decisions in each period. At the beginning of each period, the size of the batch is determined. Then, after the realizations of the random yields are observed, the various items must be allocated to customers.

It is interesting to note the similarities between these two problems. If a hotel's products are "room-nights," then it "co-produces" room-nights of luxury suites, double rooms, singles, etc. Moreover, when identical rooms are rented at different rates, their availability is a form of "co-production." The

total number of reservations that are accepted is analogous to the lotsize. The fraction of those reservations to show up as guests can be compared to the production yield. Also note that the hotel has different "grades" of room. A customer with a reservation for a given "grade" of room will gladly accept a higher grade, i.e. larger or more luxurious, room if it is offered to him at the same price.

Clearly, the two problems are not identical. The most significant differences are that the hotel faces uncertainty in demand rather than in supply, and that room-nights cannot be held in inventory. But there are enough similarities to try to use what has been learned in the manufacturing context to solve the yield management problem.

In discussions with managers of the Marriott and OMNI Hotels, it has been revealed that the reservations planning process can be broken into a long term and a short term problem. In the long term planning problem, reservations are viewed in terms of blocks of rooms. The horizon of this problem can begin several years before a target date.

In this long term planning stage, a hotel manager considers several different classes of reservations: "Airline and Government Agencies" contract for a certain number of room nights throughout a year. These customers do not specify exact dates, they are buying availability. "Corporate" reservations are made by groups who are willing to commit to specific dates and number of rooms. "Associations" negotiate a group rate for a block of rooms over a set of dates. However, they are not willing to commit themselves to rooms. They simply want the hotel to set aside an "inventory" of rooms for their association. Individual members may then call and make reservations against this inventory at the negotiated rate. Often the contracts made with these associations call for reducing the size of the block of rooms after a certain date

if too few members have reserved rooms for themselves. A special case of the Association class of reservations is "discount". If demand for rooms is expected to be low, management may set aside a block of rooms at a reduced rate. Customers who call early enough can claim reservations from this block. The final class of reservations, "Transients", is composed of individuals who pay top dollar for rooms. Although these customers usually do not make reservations until a few days before they need a room, expectations about demand from transients influences the decisions about the numbers of the other types of reservations to accept.

The short term problem covers a horizon of only 30 to 90 days prior to a target date. The manager's concerns in this problem are very different from those in the long term plan. When there are only a few weeks remaining before a target date, the blocks of reservations which were being held for association and discount groups have either been claimed by specific individuals, or they have been made available to customers in general. Some of the customers have guaranteed their reservations with a credit card. This type of reservation represents an implicit contract between the hotel and the customer in which the hotel promises to provide a room, and the customer promises to pay. Although in practice customers may not always be forced to pay if they do not show up, the "show rate" for guaranteed reservations is very high. Other customers have asked for 6 p.m. holds. This type of reservation entitles the customer to a room, as long as he arrives before 6 p.m., but costs him nothing if he does not show up. This arrangement is attractive to customers who either are uncertain as to whether they will actually need the room or are able to arrive before the 6 p.m. deadline. Naturally, the "show rate" is lower for 6 p.m. holds than for guarantees.

If all else was equal, management would prefer to have guaranteed reservations instead of 6 p.m. holds. But because requests for both new reservations and cancellations of those previously booked flow in randomly over the horizon preceding the target date, it may be desirable to accept some 6 p.m. holds to hedge against the possibility that there will be insufficient requests for guaranteed reservations to fill the hotel. On the other hand, if too many 6 p.m. holds are accepted, it may become necessary either to turn down a later request for a guarantee or, even worse, to walk a customer. Clearly, the yield management problem of balancing the inventories of these two types of reservations is a difficult one.

A number of authors have studied different versions of the short term yield management problems in airlines, hotels, rental vehicles, and a variety of other service industries. All of these problems share a common thread: uncertain demand for products which cannot be inventoried. Yet each specific application of yield management has unique features.

The airline yield management problem has been particularly well studied. Rothstein (1985) and Belobaba (1987) provide detailed reviews of mathematical approaches which have been taken to the problem of maximizing the expected revenue associated with a given flight in which capacity is fixed. Much of this work studies the allocation of sales of a single product to customers paying different fares. Belobaba (1989) has taken a marginal pricing approach in what he calls the Expected Marginal Seat Revenue (EMSR) model. In this model, he assumes that lower fare classes purchase before high ones. He shows that for each fare class i, seats should be "protected" or withheld in such a way that the marginal revenue (with respect to the number of seats withheld) from sales to higher paying classes is exactly equal to the fare for class i. Although he extends this work to account for "no shows", the emphasis

is placed upon the allocation of a fixed number of tickets to different fare categories.

Rothstein (1974) did some early work in bridging the gap between the hotel and the airline yield management problems. He contrasts the two, and proposes a Markovian sequential decision model. His focus is upon how to adjust overbooking limits at various decision points leading up to a target date. Requests for reservations, cancellations, and show rates are all sources of uncertainty.

Ladany (1976) proposes a dynamic decision model for a hotel with both single and double rooms. In each stage of his dynamic program, a random number of reservation requests and cancellations are received. The controls are the limits on the number of each type of reservation to accept. The model assumes that, on the date of the rental, no rooms are allocated until after all of the customers, both with reservations and without, have arrived. This is equivalent to assuming that all customers arrive at the same time. In practice, customers arrive throughout the day. If a room is not available when a given customer arrives, he may not be willing to wait for several hours to find out whether he will be given a room. In spite of this questionable assumption, Ladany provides a concise dynamic programming formulation of the problem. He demonstrates the application of this formulation by complete enumeration of a small problem.

Alstrup et. al. consider a similar model in an application to airlines with two types of seats. However, because of the computational effort required to solve the dynamic program for an airplane with 110 seats, they suggest that an approximation problem be solved instead. In the approximation model, passengers are treated as groups rather than individually. That is, the limits

on ticket sales can only be set at multiples of the group size. They show that it is possible to obtain accurate results using a group size of 5 or fewer.

Liberman and Yechiali(1978) propose another dynamic decision model. Although they consider only one type of room, they assume that in addition to limiting the number of reservations to accept, management can either cancel previously confirmed reservations, or acquire additional bookings at some specified cost. They show that the optimal strategy is a 3-region policy. In each period, upper and lower limits on reservation inventory create three regions. The optimal action depends only upon where the current reservation inventory lies within these regions.

Williams (1977) considers a slightly different perspective than these other authors. He models a particular date that represents a peak in demand. He assumes that demand for rooms on this date comes from three sources, listed in decreasing order of priority: (1) stayovers - guests occupying rooms on the day preceding the critical date. (2) reservations - guests arriving on the critical date with reservations. (3) walk-ins - guests arriving without reservations. He further assumes that the occupancy of the hotel on the day before the critical one is known with certainty, and calculates the expected costs of forgone revenues and overbooking that are associated with various numbers of reservations. Although Williams' suggests methods of designing decision aids, his model is concerned more with estimating the costs of specific policies than with optimization.

In the short term hotel problem, the mix of 6 p.m. holds and guaranteed reservations has a tremendous impact upon the extent to which the hotel should be "overbooked". For example, the higher the proportion of total reservations that are represented by the "low show rate" 6 p.m. holds, the more the hotel should be "overbooked". We are unaware of any previous work

that has addressed this particular aspect of the reservations problem. Although the model which we propose is conceptually similar to that of Ladany (1976), ours explicitly models the fact that the hotel must begin allocating rooms to customers before the complete arrival process has been observed.

In our investigation of this short term problem, we will focus our attention upon a "target date" T periods in the future in which the total demand for rooms on this date is likely to meet or exceed the hotel's capacity. In practice, a even the most successful hotels can expect to be filled only about four or five evenings per week. It is only for these "target dates" that the reservations problem is difficult. For other dates, when demand is expected to be less than the capacity of the hotel, the best strategy is obviously to accept all requests for reservations.

For the sake of simplicity, we will consider only one type of room. Although in practice hotels may have single rooms, double rooms, luxury suites, etc., it is not unusual for them to have most of their capacity concentrated in rooms which are indistinguishable from the customer's perspective. We will discuss an extension of our model for situations where this is not the case.

We will assume that reservations are for single night stays only. In practice, customers do make reservations for multiple (usually two or three) day stays. Hotels have even attempted to increase their utilization on low demand dates by issuing reservations for target dates only to customers requesting multiple day stays. However, customers have circumvented this policy by requesting a multiple day reservation, and subsequently canceling the dates that they do not need. Even when such gaming is not taking place, it is not unusual for a customer to show up for only a subset of the dates in his

reservation. They may either arrive late or depart early. In general, the probability that a guest will show up for a given day of his reservation does not depend upon whether he showed up the preceding day. Obviously, this assumption may be invalid in hotels that cater to long term guests. But when reservations for multi-day stays are relatively few, it is reasonable to treat them as a series of independent single night reservations.

We also assume that, during any given period, the probability that a reservation will be cancelled is independent of when it was made. The validity of this assumption was shown for airline reservations by Sanchez and Martinez (1970). Leong (1991) suggests that it is also valid for hotel reservations.

The final assumption that we make is that each reservation request is for a single room. In practice customers (tour groups, conferences, etc.) can reserve blocks of rooms. Naturally there is some dependence among the individual reservations in a block. For example, the whole group may fail to show up. However, because business travelers can be inconvenienced by large groups, many hotels accept large group reservations only on weekends when the total demand is low. Thus, it is often reasonable to assume that, on peak demand target dates, each reservation cancels or fails to show up independently.

The problem can be modeled as a series of decision points leading up to a target date. At each of these points (indexed from T, T-1, ..., 1, 0), the hotel manager sets limits on the number of each type of reservation to accept. For example, at thebeginning of the t^{th} period prior to the target date, G_{t+1} and H_{t+1} guaranteed and 6 p.m. hold reservations are held in inventory. That is, G_{t+1} and H_{t+1} reservations have already been recorded, net of any cancellations. Based upon expectations about future requests for reservations,

cancellations, and show rates, the hotel manager sets a policy, characterized by two numbers, N_{Gt} and N_{Ht} which will be followed until the beginning of period t-1. According to this policy, at most N_{Gt} and N_{Ht} new requests for guaranteed and 6 p.m. hold reservations will be accepted during period t. On the target date, decisions must be made about allocating rooms to arriving guests, including both those with reservations and walk-ins. Before presenting the formal model, we introduce the following notation:

Parameters:

π_G, π_H : The per room revenue net of variable cost for renting a room to a guaranteed or 6 p.m. hold customer.

p_G, p_H : The per room cost of failing to honor a guaranteed or 6 p.m. hold reservation.

C: The capacity of the hotel in # of rooms.

G_{T+1}, H_{T+1}: The number of guaranteed and 6 p.m. hold reservations that are outstanding at the beginning of the planning horizon.

Random Variables:

r_{Gt}: The number of requests for guaranteed reservations received during period t. Recall that the target date is period 0.

r_{Ht}: The number of requests for 6 p.m. hold reservations received during period t.

r_W: The number of room requests from "walk-in" customers on the target date.

$\mathbf{r}_t$;t=1,..,T: Two dimensional vectors whose components are r_{Gt} and r_{Ht}.

S_{Gt};t=1,..,T: The number of guaranteed reservations that "survive" period t, i.e. the number that are held at the beginning of period t and do not cancel by the end of t.

$S_{Ht};t=1,..,T$: The number of 6 p.m. hold reservations that "survive" period t, i.e. the number that are held at the beginning of period t and do not cancel by the end of t.

S_{G0}: The number of guests with guaranteed reservations to show up on the target date.

S_{H0}: The number of guests with 6 p.m. hold reservations to show up on the target date.

$\mathbf{S}_t;t=0,..,T$: A 2 dimensional vector whose components are S_{Gt} and S_{Ht}.

$G_t;t=1,..,T$: The number of guaranteed reservations that are outstanding at the end of the t^{th} period prior to the target date.

$H_t;t=1,..,T$: The number of 6 p.m. hold reservations that are outstanding at the end of the t^{th} period prior to the target date.

C_{WG} : The number of rooms that remain available for walk-ins and guaranteed reservations after the arrival of 6 p.m. hold guests.

C_G : The number of rooms that remain available for guaranteed reservations after the arrival of walk-ins and 6 p.m. hold guests.

Decision Variables:

$N_{Gt};t=1,..,T$: The limit on the number of guaranteed reservations to accept during the t^{th} period prior to the target date.

$N_{Ht};t=1,..,T$: The limit on the number of 6 p.m. hold reservations to accept during the t^{th} period prior to the target date.

N_{HA} : The number of rooms that are assigned to customers arriving with 6 p.m. hold reservations on the target date.

N_{HW} : The number customers with 6 p.m. hold reservations who arrive on the target date, but do not receive a room.

N_W: The limit on the total number of walk-in requests that will be accepted on the target date.

N_{GA} : The number of rooms that are assigned to customers arriving with guaranteed reservations on the target date.

N_{GW} : The number customers with guaranteed reservations who arrive on the target date, but do not receive a room.

Let us also define the notation $f_{\mathbf{X}}(\mathbf{x})$ as the probability mass function (or, more generally, the probability density function) for the random vector $\mathbf{X}$ evaluated at the point $\mathbf{x}$. Consider the following model:

$$MER_t(G_{t+1},H_{t+1},C) = \underset{N_{Gt},N_{Ht}}{Max}\{ER_t(G_{t+1},H_{t+1},C,N_{Gt},N_{Ht})\} =$$

$$\underset{N_{Gt},N_{Ht}}{Max}\left\{\sum_{r_{Gt}=0}^{\infty}\sum_{r_{Ht}=0}^{\infty}\sum_{S_{Gt}=0}^{G_{t+1}}\sum_{S_{Ht}=0}^{H_{t+1}} MER_{t-1}(G_t,H_t,C)f_{\mathbf{r},\mathbf{S}}(\mathbf{r}_t,\mathbf{S}_t)\right\},\ \text{for } t = T,\ldots,1 \qquad 7.1$$

where,

$$H_t = Min\{r_{H,t}, N_{H,t}\} + S_{H,t},\ \text{for } t = 1,\ldots,T \qquad 7.2$$

$$G_t = Min\{r_{G,t}, N_{G,t}\} + S_{G,t},\ \text{for } t = 1,\ldots,T \qquad 7.3$$

$$MER_0(G_1,H_1,C) =$$

$$\underset{\substack{N_{H0}\le C\\ N_{H0}\le H_1}}{Max}\{ER_H(G_1,H_1,C,N_{H0})\} = \underset{\substack{N_{H0}\le C\\ N_{H0}\le H_1}}{Max}\left\{\sum_{S_{H0}=0}^{H_1} R_H(G_1,S_{H0},C,N_{H0})f_{H_0}(H_0)\right\} \qquad 7.4$$

where:

$$R_H(G_1,S_{H0},C,N_{H0}) = \pi_H N_{HA} - p_H N_{HW} + MER_W(G_1,C-N_{HA}) \qquad 7.5a$$

$$\text{s.t.:}\quad N_{HA} = Min(N_{H0}, S_{H0}) \qquad 7.5b$$

$$N_{HA} + N_{HW} = S_{H0} \qquad 7.5c$$

$$MER_W(G_1,C_{WG}) = \underset{N_W\le C_{WG}}{Max}\{ER_W(G_1,C_{WG},N_W)\}$$

$$= \underset{N_W\le C_{WG}}{Max}\left\{\sum_{r_W=0}^{\infty} R_W(G_1,r_W,C_{WG},N_W)f_{r_W}(r_W)\right\} \qquad 7.6$$

where:

$$R_W(G_1, r_W, C_{WG}, N_W) = \pi_W W + ER_G(G_1, C_{WG} - W) \quad 7.7a$$

$$\text{s.t.:} \quad W = \text{Min}(N_W, r_W) \quad 7.7b$$

$$ER_G(G_1, C_G) = \left\{ \sum_{S_{G0}=0}^{G_1} R_G(S_{G0}, C_G) f_{S_{G0}}(S_{G0}) \right\} \quad 7.8$$

where:

$$R_G(S_{G0}, C_G) = \text{Max:} \ \pi_G N_{GA} - p_G N_{GW} \quad 7.9a$$

$$\text{s.t.:} \ N_{GA} \leq C_G \quad 7.9b$$

$$N_{GA} + N_{GW} = S_{G0} \quad 7.9c$$

$$N_{GA}, N_{GW} \geq 0 \quad 7.9d$$

As shown in Figure 7.1, this model has the following interpretation: At the beginning of each of the t= 1,...,T periods preceding the target date, the manager has inventories of G_{t+1} and H_{t+1} guaranteed and 6 p.m. hold reservations respectively. He determines a policy which places limits N_{Gt} and N_{Ht} on the numbers of reservations that he will accept prior to the next decision point. $S_{G,t-1}$ and $S_{H,t-1}$ represent the "surviving reservations", i.e. the number that were held at the beginning of period t and did not cancel by the end. Note that the probability distribution of S_{Gt} (S_{Ht}), the number of guaranteed (6 p.m. hold) reservations which survive until t, depends on G_{t+1} (H_{t+1}), the number that were held at the the end of t+1 (or, equivalently, beginning of t). The actual size of the reservation inventories at the end of t (beginning of t-1) are functions of the management imposed limits (N_{Gt} and N_{Ht}), the number of surviving reservations (S_{Gt} and S_{Ht}), and the number of requests for new reservations (r_{Gt} and r_{Ht}). For example, suppose that t periods before a target date there are G_{t+1} outstanding guaranteed reservations, and that management decides that it will accept a maximum of N_{Gt} guaranteed reservations during period t. The total inventory of

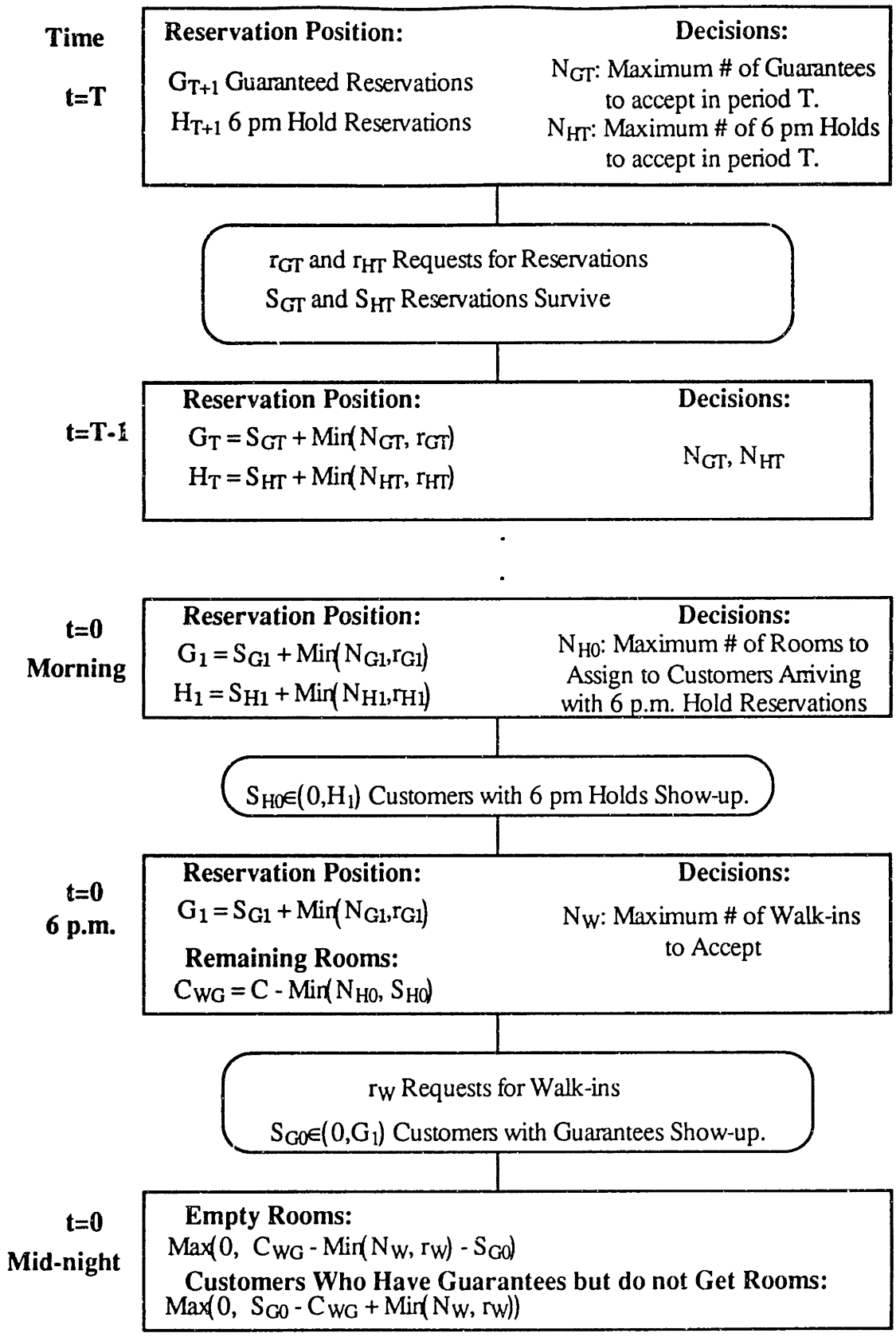


Figure 7.1: Reservations Process

reservations at the beginning of period t-1 is at least as large as S_{Gt}, the number of reservations surviving from period t. In addition, the number of newly accepted reservations is either the limit N_{Gt}, or the number of requests made in period t, r_{Gt}.

Note that, this model assumes that the limit on the number of reservations to accept cannot be adjusted between decision points. For example, suppose that during period t, the number of requests for guaranteed reservations quickly exceeds the limit N_{Gt}. Then regardless of how many cancellations occur during, period t, no more new bookings would be accepted. Our model approximates what occurs in practice where limits are imposed upon net inventories of reservations instead of upon new bookings. However, it should be noted that by using decision periods that are sufficiently short, this approximation can be made arbitrarily good.

At the beginning of the target day itself (t=0), the hotel manager faces a problem that is slightly different from the ones in earlier decision periods. Here, he is dealing with the arrival of customers with and without reservations, i.e. walk-ins. Problem $MER_0(G_1,H_1,C)$ represents the maximum expected profit given that the hotel has a capacity of C rooms, and that G_1 and H_1 reservations are held in the morning of the target date. The actual profits are determined by the show rates of these reservations, the number of walk-in customers, and the hotel manager's decisions about room assignments.

We have assumed that the three different types of guests arrive at non-overlapping times of the day. In practice, the latest arriving customers tend to have guaranteed reservations. It is also generally the case that, on days when the hotel is likely to be filled, management accepts walk-ins only after the expiration of the 6 p.m. holds. In other cases, when the hotel is

not likely to be filled, the solution to the problem is trivial: all arriving customers are assigned rooms. Thus, for the difficult cases, when capacity of the hotel is constrained there is little overlap between the arrival of the 6 p.m. hold customers and walk-ins. When some of the customers with guaranteed reservations arrive prior to the end of the arrivals of the other two groups, we need only modify the way in which we use the function: $ER_W(G_1,C_{WG},N_W)$. The parameter G_1 should reflect the number of outstanding guaranteed reservations, i.e. net of early arrivals, at the time the walk-in problem is solved. The parameter C_{WG} should reflect the number of rooms that are still available after the arrival of 6 p.m. holds and early arrivals of customers with guaranteed reservations.

The first decision to be made on the target date is N_{H0}, the limit on the number of rooms to assign to customers with 6 p.m. holds. The actual number that are assigned depends on this limit as well as on the number (S_{H0}) of customers who arrive with valid 6 p.m. hold reservations. In many cases it is less expensive, to "walk" a customer at 6 p.m. than to do so later in the evening. Not only is it easier to find an alternative hotel at 6 p.m., the customer's perception of the inconvenience of being walked is likely to increase with the time of evening. Because customers with guaranteed reservations often arrive very late in the evening, it may be prudent to "walk" some customers who have 6 p.m. hold reservations in order to serve these late arrivals.

The next decision occurs after the 6 p.m. hold customers have either arrived or their reservations have expired. During the period between 6 p.m. and the end of the evening, the hotel manager may receive requests for rooms from walk-in customers. Because these customers do not have reservations, there is no penalty, other than foregone profit margin for

turning them away. $MER_W(G_1, C_{WG})$ represents the decision regarding N_W, the limit on the number of rooms to give to walk-in customers. The function $R_W(G_1,r_W,C_{WG},N_W)$ represents the expected profits from the arrivals of walk-ins and customers with guaranteed reservations, given that:

1) C_{WG} rooms remain available after the arrival of the 6 p.m. hold guests.
2) r_W walk-in customers request rooms.
3) The limit on the number of walk-ins to be given rooms is N_W.

Notice that N_W is constrained to be no larger than the number of available rooms.

Finally, the function $ER_G(G_1,C_G)$ represents the expected revenue from the arrival of guaranteed reservations given that:

1) C_G rooms remain available after the arrival of 6 p.m. hold guests and walk-ins.
2) G_1 guaranteed reservations are booked.

The expectation is taken with respect to S_{G0}, the actual number of guests with guaranteed reservations to arrive expecting rooms.

Note that the model allows for different prices for 6 p.m. holds, guaranteed reservations, and walk-ins, but does not allow for these prices to vary over time. In practice, hotels often offer discounts in order to encourage customers to book reservations in advance. If demand for reservations is sufficiently high, as the target date approaches, these discounts may be reduced or eliminated. Thus the last customers to book reservations may pay the same high price as walk-ins. Although our

model assumes that the prices are time invariant, we discuss an extension to time varing prices in section 12.

In general, the probability distributions for reservation requests and walk-ins can best be modeled by Poisson distributions. Those for the numbers of reservations to "survive" from one period to the next are best modeled as Binomials. However, these discrete distributions are difficult to work with. Because most hotels have several hundred rooms, little accuracy is lost by using continuous distributions as approximations. When it is reasonable to use continuous approximations, the preceding model can be re-stated as follows:

$$MER_t(G_{t+1},H_{t+1},C) =$$

$$\underset{N_{Gt},N_{Ht}}{Max}\{ER_t(G_{t+1},H_{t+1},C,N_{Gt},N_{Ht})\} =$$

$$\underset{N_{Gt},N_{Ht}}{Max}\left\{\int_{r_t=0}^{\infty}\int_{s_t=0}^{1} MER_{t-1}(G_t,H_t,C) f_{r_t,S_t}(r_t,S_t)\,dr_t\,dS_t\right\} \quad \text{for } t = T,\ldots, 1 \qquad 7.10$$

where:

$$H_t = Min\{r_{H,t}, N_{Ht}\} + S_{Ht}, \text{ for } t = 1,\ldots,T \qquad 7.11a$$

$$G_t = Min\{r_{G,t}, N_{Gt}\} + S_{Gt}, \text{ for } t = 1,\ldots,T \qquad 7.11b$$

$$MER_0(G_1,H_1,C) = \underset{\substack{N_{H0}\le C\\ N_{H0}\le H_1}}{Max}\{ER_H(G_1,S_{H0},C,N_{H0})\}$$

$$= \underset{\substack{N_{H0}\le C\\ N_{H0}\le H_1}}{Max}\left\{\int_{S_{H0}=0}^{H_1} R_H(G_1,S_{H0},C,N_{H0}) f_{S_{H0}}(S_{H0})\,dS_{H0}\right\} \qquad 7.12$$

where, as in the dicrete version of the model:

$$R_H(G_1,S_{H0},C,N_{H0}) = \pi_H N_{HA} - p_H N_{HW} + MER_W(G_1,C-N_{HA}) \quad 7.5a$$

$$\text{s.t.:} \quad N_{HA} = \text{Min}(N_{H0}, S_{H0}) \quad 7.5b$$

$$N_{HA} + N_{HW} = S_{H0} \quad 7.5c$$

$$MER_W(G_1,C_{WG}) = \max_{N_W \le C_{WG}} \{ER_W(G_1,C_{WG},N_W)\}$$

$$= \max_{N_W \le C_{WG}} \left\{ \int_{r_W=0}^{\infty} R_W(G_1,r_W,C_{WG},N_W) f_{r_W}(r_W) dr \right\} \quad 7.13$$

where, as in the discrete version of the model:

$$R_W(G_1,r_W,C_{WG},N_W) = \pi_W W + ER_G(G_1,C_{WG}-W) \quad 7.7a$$

$$\text{s.t.:} \quad W = \text{Min}(N_W, r_W) \quad 7.7b$$

$$ER_G(G_1,C_G) = \left\{ \int_{S_{G0}=0}^{G_1} R_G(S_{G0},C_G) f_{S_{G0}}(S_{G0}) dS_{G0} \right\} \quad 7.14$$

where, as in the discrete version of the model:

$$R_G(S_{G0}, C_G) = \quad \text{Max: } \pi_G N_{GA} - p_G N_{GW} \quad 7.9a$$

$$\text{s.t.:} \quad N_{GA} \le C_G \quad 7.9b$$

$$N_{GA} + N_{GW} = S_{G0} \quad 7.9c$$

$$N_{GA}, N_{GW} \ge 0 \quad 7.9d$$

In Section 8, we analyze the mathematics of the formulation above. In Chapter 9 we use the the insights from this analysis and our observations of what is done in practice to develop heuristic methods for solving the reservations problem. In Chapter 10, we describe an upper bound on the value of an optimal solution to the dynamic programming formulation of the problem. Finally, in Chapters 11 and 12, we discuss the performance of our heuristics and areas for future research.

Chapter 8: Model Analysis

The dynamic programming formulation in $MER_t(G_{t+1}, H_{t+1})$ is not an easy one to solve. In many practical instances, a single hotel may have several hundred rooms, and the state space becomes quite large. Although Ladany (1976) discusses the performance of a total enumeration algorithm for a similar reservation problem, this approach may not be practical for managers of large hotels to use on a daily basis.

Rather than attempting to enumerate the discrete random variable formulation of $MER_t(G_{t+1}, H_{t+1})$, we have focused our attention on the continuous random variable formulation. Hotel managers are not often willing to commit themselves to a single set of parameters for the probability distributions of reservation requests and survival rates, and may want to explore several different scenarios before making a decision. The objective of our analysis is to provide insight into the underlying trade-offs that are involved in the acceptance of guaranteed and 6 p.m. hold reservations and walk-ins.

Before discussing optimality conditions we will investigate the concavity of the problem. We will first show that, the expected revenue (ER) associated with the arrival of customers with guaranteed reservations is a concave function of the remaining capacity, i.e. the number of rooms which have not yet been allocated.

Lemma 8.1: $ER_G(G_1,C_G)$ is non-decreasing and concave in C_G.

Proof: The function $R_G(S_{G0},C_G)$ is the solution of a maximization linear program in which C_G is a constant parameter on the right hand side of a "less than or equal to" constraint. Thus $R_G(S_{G0},C_G)$ is concave in C_G, and, because a convex combination of concave functions is also concave,

$ER_G(G_1,C_G)$ is concave in C_G. To see that $ER_G(G_1,C_G)$ is non-decreasing in C_G, we observe that for a given realization of S_{G0}, constraint 7.9 of the linear program $R_G(S_{G0},C_G)$ is less restrictive as C_G increases. QED

Lemma 8.2: The gradient of $ER_G(G_1,C_G)$ with respect to C_G is as follows:

$$\frac{\partial}{\partial C_G}[ER_G(G_1,C_G)] = (\pi_G + p_G)F^c_{S_{G0}}(C_G) \qquad 8.1$$

where $F^c_{S_{G0}}(C_G)$ is the probability that $S_{G0} \geq C_G$.

Proof: By inspection of linear program $R_G(S_{G0},C_G)$, shown in equations 7.9a-d, it can be seen that C_G appears only in constraint 7.9b, and that constraint is binding only for realizations of S_{G0} that are greater than C_G. It is also obvious from inspection that when constraint 7.9b is binding, increasing C_G by δ will cause the optimal solution vector $(N_{GA},N_{GW})^*$ to change by $(\delta,-\delta)$, and the value of the optimal solution to change by $\delta(\pi_G+p_G)$. Thus, the shadow price of constraint 7.9b is π_G+p_G. It follows that the gradient of the expected value of $R_G(S_{G0},C_G,)$ is equal to this shadow price multiplied by the probability that constraint 7.9b is binding. QED

The above result is consistent with the intuition that the expected profit associated with the arrival of customers with guaranteed reservations increases as a function of C_G, the number of rooms that are available for them. However, as C_G increases, the marginal value of an additional room diminishes until it reaches zero when $C_G = G_1$. In other words, there is nothing to be gained from having more rooms than there are customers to claim them.

We are now prepared to begin discussing the optimality conditions for each of the various stages of the dynammic program.

Claim 8.1: An optimal solution to problem $ER_W(G_1,C_{WG},N_W)$ can be found at the point:

$$N_W^*(C_{WG}) = \begin{cases} 0 & \text{if } C_{WG} \le F_{S_{G0}}^{c\,-1}\left(\dfrac{\pi_W}{\pi_G + p_G}\right) \\ C_{WG} - F_{S_{G0}}^{c\,-1}\left(\dfrac{\pi_W}{\pi_G + p_G}\right) & \text{otherwise} \end{cases} \qquad 8.2$$

where $F_{S_{G0}}^{c\,-1}(\alpha) = G : \text{Prob}(S_{G0} \ge G) = \alpha$.

Proof: It can be shown that $ER_W(G_1,C_{WG},N_W)$ is pseudoconcave. The proof of this is given in appendix 8A. Thus, the following is a sufficient condition for N_W^* to be a global maximum (Mangasarian, p. 145):

$$(N_W^* + \Delta)\cdot \frac{\partial}{\partial N_W}[ER_W(G_1,C_{WG},N_W^*)] \le 0, \qquad 8.3$$

for any feasible direction Δ.

Before attempting to compute the gradient of $R_W(G_1,r_W,C_{WG},N_W)$, we observe that the function in equations 7.7a-b is equivalent to:

$$R_W(G_1,r_W,C_{WG},N_W) = \pi_W r_W + ER_G(G_1,C_{WG}-r_W) \quad \text{for } r_W \le N_W \qquad 8.4a$$

$$= \pi_W N_W + ER_G(G_1,C_{WG}-N_W) \quad \text{for } r_W > N_W \qquad 8.4b$$

To show that condition 8.3 is satisfied at our proposed point, we have:

$$\frac{\partial}{\partial N_W}[ER_W(G_1,r_W,C_{WG},N_W)] = \frac{\partial}{\partial N_W}\left[\int_{r_W=0}^{\infty} R_W(G_1,r_W,C_{WG},N_W) f_r(r_W)dr\right]$$

$$= \int_{r_W=N_W}^{\infty} \left(\pi_W + \frac{\partial}{\partial N_W}[ER_G(G_1,C_{WG}-N_W)]\right) f_r(r_W)dr \qquad 8.5$$

Substituting for the partial derivative of $ER_G(G_1,C_{WG}-N_W)$ with respect to $(C_{WG}-N_W)$ and applying the chain rule:

$$\frac{\partial}{\partial N_W}[ER_W(G_1,r_W,C_{WG},N_W)] =$$

$$= \int_{r_W=N_W}^{\infty} \left(\pi_W - (\pi_G+p_G)\left(F^c_{S_{G0}}(C_{WG}-N_W)\right)\right) f_r(r_W)dr \qquad 8.6$$

$$= \left[\pi_W - (\pi_G+p_G)\left(F^c_{S_{G0}}(C_{WG}-N_W)\right)\right] \cdot F^c_{r_W}(N_W) \qquad 8.7$$

We can now substitute expressions 8.2 and 8.7 into the optimality condition for a pseudo-concave function which is given in equation 8.3. There are two cases:

CASE 1: $C_{WG} < F^{c\,-1}_{S_{G0}}\left(\frac{\pi_W}{\pi_G + p_G}\right)$: from 8.2, $N^*_W(C_{WG}) = 0$, and we have:

$$\left(N^*_W(C_{WG}) + \Delta\right) \cdot \frac{\partial}{\partial N_W}\left[\int_{r_W=0}^{\infty} R_W(G_1,r_W,C_{WG},N_W) f_r(r_W)dr\right]$$

$$= [0+\Delta]\left[\pi_W - (\pi_G+p_G)\left(F^c_{S_{G0}}(C_{WG})\right)\right]\left[F^c_{r_W}(0)\right] < 0 \quad \text{for all } \Delta > 0. \qquad 8.8$$

where, because the complementary cummulative distribution is a decreasing function: $\frac{\pi_W}{\pi_G+p_G} < F^c_{S_{G0}}(C_{WG}) \leftrightarrow F^{c\,-1}_{S_{G0}}\left(\frac{\pi_W}{\pi_G+p_G}\right) > C_{WG}$.

Note that when $N_W^*(C_{WG}) = 0$, a feasible direction Δ must be > 0.

CASE 2: $C_{WG} \geq F^c_{S_{G0}}{}^{-1}\left(\frac{\pi_W}{\pi_G + p_G}\right)$: From 8.2, $N^*_W(C_{WG}) = C_{WG} - F^c_{S_{G0}}{}^{-1}\left(\frac{\pi_W}{\pi_G+p_G}\right)$,

and we have:

$$\left(N^*_W(C_{WG}) + \Delta\right) \cdot \frac{\partial}{\partial N_W}\left[\int_{r_W=0}^{\infty} R_W(G_1, r_W, C_{WG}, N_W) f_r(r_W) dr\right]$$

$$= \left[N^*_W(C_{WG})+\Delta\right]\left[\frac{\pi_W}{\pi_G+p_G} - \left(F^c_{S_{G0}}\left(C_{WG}-N^*_W(C_{WG})\right)\right)\right]\left[F^c_{r_W}\left(N^*_W(C_{WG})\right)\right] = 0$$

for all Δ. 8.9

QED

Let us consider the intuition behind the the optimal solution to MER_W:

$$N^*_W: \quad \frac{\pi_W}{\pi_G+p_G} = F^c_{S_{G0}}\left(C_{WG}-N^*_W\right) = \left(1 - F_{S_{G0}}\left(C_{WG}-N^*_W\right)\right).$$

This bears a striking resemblance to the optimality condition for the classic newsboy problem. In that problem, it is optimal to order a quantity of newspapers such that the probability of a stockout equals the ratio of backorder costs to the sum of backorder and holding costs. Here, rooms should be allocated to walk-in requests until the probability of failing to provide a room to a guaranteed reservation exceeds the ratio: $\pi_w /(\pi_G+p_G)$. Recall that $C_G = C_{WG}-N_W$ is the number of rooms that are available for guaranteed reservations if N_W rooms are given to walk-ins.

In Figure 8.1, we can see that, for a given number of guaranteed reservations, there is an optimal number of rooms which should be saved for these guests. Walk-ins should be accepted until the number of

remaining rooms equals this critical value. If the number of rooms remaining after the arrival of the 6 p.m. hold guests is less than this critical value, then no walk-ins should be accepted.

We can now investigate the optimality conditions for the problem of deciding the number of rooms to allocate to customers who arrive with 6 p.m. hold reservations, i.e. problem $MER_0(G_1,H_1)$.

Optimal number of walk-ins as a function of the number of rooms remaining after the 6 p.m. hold customers arrive.

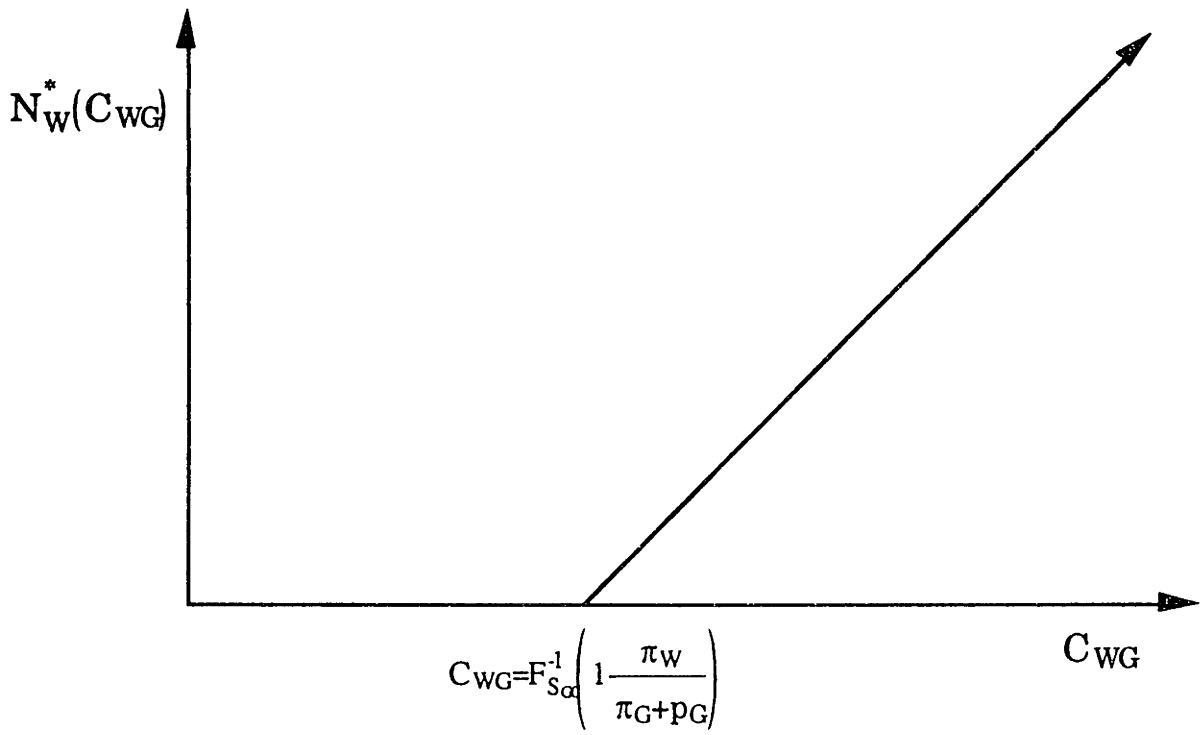


Figure 8.1

Claim 8.2: If $\pi_W < \pi_H+p_H$, then the function $ER_H(G_1,S_{H0},C,N_{H0})$ is maximized with respect to N_{H0} at the point:

$$N_{H0}^{*} = \begin{cases} 0 & \text{if } \pi_H + p_H - (\pi_G+p_G)F^{c}_{S_{G0}}(C) < 0 \\ C - F^{c}_{S_{G0}}{}^{-1}\left(\dfrac{\pi_H+p_H}{\pi_G+p_G}\right) & \text{otherwise} \end{cases} \qquad 8.10$$

Proof: It can be shown that $ER_H(G_1,S_{H0},C,N_{H0})$ is a pseudo-concave functionof N_{H0}. The details of this proof are provided in appendix 8B. Thus, the following is a sufficient condition for $ER_H(G_1,S_{H0},C,N_{H0})$ to attain a global maximum at $N_{H0} = N_H^*$:

$$(N_H^* + \Delta)\cdot \frac{\partial}{\partial N_{H0}}[ER_H(G_1,S_{H0},C,N_H^*)] \le 0, \text{ for any feasible direction } \Delta. \qquad 8.11$$

As shown in appendix 8B, the gradient of $ER_H(G_1,S_{H0},C,N_{H0})$ with respect to N_{H0} can be expressed as the following:

CASE 1: $N_{H0} < C\text{-}F^{c}_{S_{G0}}{}^{-1}\left(\dfrac{\pi_W}{\pi_G+p_G}\right)$:

$$\frac{\partial}{\partial N_{H0}}[ER_H(G_1,S_{H0},C,N_{H0})] =$$

$$\left(\pi_H + p_H - \pi_W F^{c}_{r_W}\left(C\text{-}N_{H0}\text{-}F^{c}_{S_{G0}}{}^{-1}\left(\frac{\pi_W}{\pi_G+p_G}\right)\right)\right)F^{c}_{S_{H0}}(N_{H0})$$

$$-(\pi_G+p_G)\int_{r_W=0}^{C\text{-}N_{H0}\text{-}F^{c}_{S_{G0}}{}^{-1}\left(\frac{\pi_W}{\pi_G+p_G}\right)} F_{S^{c}_{G0}}(C\text{-}N_{H_0}\text{-}r_W) f_{r}(r_W)dr \ge \pi_H + p_H - \pi_W \qquad 8B.8a$$

CASE 2: $N_{H0} \ge C\text{-}F^{c}_{S_{G0}}{}^{-1}\left(\dfrac{\pi_W}{\pi_G+p_G}\right)$:

$$\frac{\partial}{\partial N_{H0}}[ER_H(G_1,S_{H0},C,N_{H0})] = \left(\pi_H+p_H - (\pi_G+p_G)F^{c}_{S_{G0}}(C\text{-}N_{H0})\right)F^{c}_{S_{H0}}(N_{H0}) \qquad 8B.8b$$

Suppose that: $C \le F^{c}_{S_{\infty}}{}^{-1}(\pi_W/(\pi_G+p_G))$, hence $F^{c}_{S_{\infty}}(C) \ge \pi_W/(\pi_G+p_G)$. This condition implies that, regardless of how many rooms are assigned to 6 p.m. hold customers, the optimal walk-in policy (claim 8.1) is to accept no walk-ins. It also implies that for all $N_{H0} \ge 0$, 8B.8b is the relevent gradient of $ER_H(G_1,S_{H0},C,N_{H0})$. Thus, because $F^{c}_{S_{\infty}}(C-N_{H0})$ is increasing in N_{H0}, it follows that, if $\pi_H + p_H - (\pi_G+p_G)F^{c}_{S_{\infty}}(C) \le 0$, then expression 8B.9 is less than or equal to zero for all $N_{H0} \ge 0$, and the optimal solution is $N_{H0}^{*} = 0$.

Now suppose that: $C > F^{c}_{S_{\infty}}{}^{-1}(\pi_W/(\pi_G+p_G))$, hence $F^{c}_{S_{\infty}}(C) < \pi_W/(\pi_G+p_G)$. This condition implies that the number of rooms that will be allocated to walk-ins depends upon the number that are allocated to 6 p.m. holds. As noted in appendix 8B, the gradient of $ER_H(G_1,S_{H0},C,N_{H0})$ with respect to N_{H0} is positive for $N_{H0} < C - F^{c}_{S_{\infty}}{}^{-1}(\pi_W/(\pi_G+p_G))$, and is non-increasing for $N_{H0} \ge C - F^{c}_{S_{\infty}}{}^{-1}(\pi_W/(\pi_G+p_G))$. It follows that the optimality condition in equation 8.11 is satisfied at the point:

$$N^{*}_{H0} = C - F^{c}_{S_{\infty}}{}^{-1}\left(\frac{\pi_H+p_H}{\pi_G+p_G}\right)$$

QED

The intuition of this gradient is the following: Suppose that the capacity of the hotel is large enough to accomodate all of the customers with guaranteed reservations. As we begin to allocate rooms to 6 pm hold customers, we trade off the increased revenue and savings in overbooking penalties associated with them against the potential lost revenue from turning away walk-ins. Note that the final term in 8B.8a represents the expected benefit associated with rooms which we would like to sell to walk-ins but are unable to because of limited demand. However, as we continue

to increase N_{H0}, we decrease the amount of capacity that is available to walk-ins and guarantees to the point at which no walk-ins will be accepted. When this occurs the increased revenue and savings in "overbooking" penalties from allocating an additional room to a 6 pm hold is balanced against the potential lost revenue and overbooking penalty of turning away a guarantee.

As shown in the preceding discussion, the room allocation problem on the target day lends itself to mathematical analysis. Under only minor restrictions, it is possible to derive closed form, optimal solutions to both the 6 p.m. hold and the walk-in allocation problems. Unfortunately, the reservations acceptance problems that must be solved prior to the target date are much more difficult to analyze. The inherent difficulty in these problems arises from the fact that the functional forms of the probability distributions for state variables G_t and H_t depend upon decisions made in periods T, T-1, ..., t. In particular, the functional form of the probability distribution of cancellations (or no-shows) during period t depends upon the limits that were placed upon accepting reservations in periods prior to t.

Rather than attempting to solve the reservations acceptance problem analytically or via enumeration, we have focused upon developing heuristics which can be shown to provide solutions that are near optimal. In the next section, we describe three heuristics which were developed on the basis of discussions we held with hotel managers as well as the insight that we have gained from analyzing the optimal solutions to the room allocation problems.

Appendix 8A

Claim: $ER_W(G_1, r_W, C_{WG}, N_W)$ is a pseudoconcave function of N_W.

Proof: In order to prove the claim, it suffices to show that:

For any $N_1, N_2 \in \left(0, r_W^{max}\right)$, where $r_W^{max} = \sup\{r_W : f_r(r_W) > 0\}$:

$\nabla_N ER_W(G_1, r_W, C_{WG}, N_1) \cdot (N_2 - N_1) \leq 0$ implies that:

$$ER_W(G_1, r_W, C_{WG}, N_1) \geq ER_W(G_1, r_W, C_{WG}, N_2) \qquad 8A.1$$

A more general definition of pseudo concavity can be found in Mangasarian (1969). We will show this by demonstrating that the following set:

$$\Gamma = \{N : N \geq 0 \text{ and } \nabla_N [ER_W(G_1, r_W, C_{WG}, N)] > 0\}$$

is convex. Consider the partial derivative of $ER_W(G_1, r_W, C_{WG}, N_W)$ with respect to N_W. Recall from 8.7:

$$\nabla_{N_W}[ER_W(G_1, r_W, C_{WG}, N_W)] = \frac{\partial}{\partial N_W}[ER_W(G_1, r_W, C_{WG}, N_W)] =$$

$$= \left[\pi_W - (\pi_G + p_G) F^c_{S_{G0}}(C_{WG} - N_W)\right] \cdot F^c_{r_W}(N_W) \qquad 8.7$$

This gradient is positive if and only if:

$$\frac{\pi_W}{\pi_G + p_G} > F^c_{S_{G0}}(C_{WG} - N_W), \text{ and } F^c_{r_W}(N_W) > 0. \qquad 8A.2$$

Suppose that the gradient is positive and this condition is satisfied at two points: $N_W^1 \geq 0$ and $N_W^2 \geq 0$, where $N_W^1 \neq N_W^2$. Then for any $\lambda \in (0,1)$:

$$\frac{\pi_W}{\pi_G + p_G} > F^c_{S_{G0}}\left(C_{WG} - (\lambda N_W^1 + (1-\lambda) N_W^2)\right), \text{ and } F^c_{r_W}(\lambda N_W^1 + (1-\lambda) N_W^2) > 0. \qquad 8A.3$$

Thus, the gradient is also positive at $N_W = \lambda N_W^1 + (1-\lambda) N_W^2$. It follows that the set $\Gamma = \{N : N \geq 0 \text{ and } \nabla_N [ER_W(G_1, r_W, C_{WG}, N)] > 0\}$ is convex.

QED

Appendix 8B

Claim: If $\pi_W < \pi_H + p_H$, then $ER_H(G_1, S_{H0}, C, N_{H0})$ is a pseudo-concave function of N_{H0}.

Proof: To show that $ER_H(G_1, S_{H0}, C, N_{H0})$ is pseudo-concave, we must establish that:

For any $N_1, N_2 \in (0, \text{Min}(C, N_{H1}))$,

$\nabla_N [ER_H(G_1, S_{H0}, C, N_1)][N_2 - N_1] \leq 0$ implies that:

$ER_H(G_1, S_{H0}, C, N_1) \geq ER_H(G_1, S_{H0}, C, N_2)$.

We will show this by demonstrating that the following set:

$$\Gamma = \{N: N \geq 0 \text{ and } \nabla_N [ER_H(G_1, S_{H0}, C, N)] > 0\}$$

is convex. Consider the partial derivative of $ER_H(G_1, S_{H0}, C, N_{H0})$ with respect to N_{H0} :

$$\frac{\partial}{\partial N_{H0}}[ER_H(G_1, S_{H0}, C, N_{H0})] = \frac{\partial}{\partial N_{H0}}\left[\int_{S_{H0}=0}^{H_1} R_H(G_1, S_{H0}, C, N_{H0}) f_{S_{H0}}(S_{H0}) dS_{H0}\right]$$

$$= \frac{\partial}{\partial N_{H0}}\left[\int_{S_{H0}=0}^{N_H} (\pi_H S_{H0} + MER_W(G_1, C - S_{H0})) f_{S_{H0}}(S_{H0}) dS_{H0}\right]$$

$$+ \frac{\partial}{\partial N_{H0}}\left[\int_{S_{H0}=N_H}^{H_1} ((\pi_H + p_H) N_H - p_H S_{H0} + MER_W(G_1, C - N_H)) f_{S_{H0}}(S_{H0}) dS_{H0}\right] \quad \text{8B.1}$$

$$= \left[\int_{S_{H0}=N_H}^{H_1} \left(\pi_H + p_H + \frac{\partial(C - N_{H0})}{\partial N_{H0}} \frac{\partial(MER_W(G_1, C - N_{H0}))}{\partial(C - N_{H0})}\right) f_{S_{H0}}(S_{H0}) dS_{H0}\right] \quad \text{8B.2}$$

$$= \left[\int_{S_{H0}=N_{H0}}^{H_1} \left(\pi_H + p_H - \frac{\partial(MER_W(G_1, C-N_{H0}))}{\partial(C-N_{H0})} \right) f_{S_{H0}}(S_{H0}) dS_{H0} \right] \qquad 8B.3$$

Further analysis of expression 8B.3 necessitates that we evaluate the partial derivative of $MER_W(G_1,C)$ with respect to C:

$$\frac{\partial}{\partial C} MER_W(G_1,C) = \frac{\partial}{\partial C} \left[\int_{r_W=0}^{N_W^*(C)} (\pi_W r_W + ER_G(G_1, C-r_W)) f_r(r_W) dr \right]$$

$$+ \frac{\partial}{\partial C} \left[\int_{r_W=N_W^*(C)}^{\infty} (\pi_W N_W^*(C) + ER_G(G_1, C-N_W^*(C))) f_r(r_W) dr \right] \qquad 8B.4$$

$$= \int_{r_W=0}^{N_W^*(C)} \frac{\partial}{\partial C} (\pi_W r_W + ER_G(G_1, C-r_W)) f_r(r_W) dr$$

$$+ \int_{r_W=N_W^*(C)}^{\infty} \left(\pi_W \frac{\partial N_W^*(C)}{\partial C} + \frac{\partial ER_G(G_1, C-N_W^*(C))}{\partial (C-N_W^*(C))} \frac{\partial (C-N_W^*(C))}{\partial C} \right) f_r(r_W) dr \qquad 8B.5$$

$$= (\pi_G + p_G) \int_{r_W=0}^{N_W^*(C)} F_{S_{G0}}^c(C-r_W) f_r(r_W) dr$$

$$+ \left(\pi_W \frac{\partial N_W^*(C)}{\partial C} + (\pi_G + p_G) F_{S_{G0}}^c(C-N_W^*(C)) \frac{\partial (C-N_W^*(C))}{\partial C} \right) F_{r_W}^c(N_W^*(C)) \qquad 8B.6$$

Expression 8B.6 is obtained from 8B.5 by substituting expression 8.1 for the partial derivative of ER_G and integrating with respect to r_W. From the definition of $N_W^*(C)$ in Claim 8.1 it is easy to see that :

$$\text{If } C < F^{c}_{S_{G0}}{}^{-1}\left(\frac{\pi_W}{\pi_G+p_G}\right): N^{*}_{W}(C) = 0,\ \frac{\partial N^{*}_{W}(C)}{\partial C} = 0,\ \text{and } \frac{\partial\left(C-N^{*}_{W}(C)\right)}{\partial C} = 1.$$

$$\text{Otherwise: } N^{*}_{W}(C) = C - F^{c}_{S_{G0}}{}^{-1}\left(\frac{\pi_W}{\pi_G+p_G}\right),\ \frac{\partial N^{*}_{W}(C)}{\partial C} = 1, \text{ and } \frac{\partial\left(C-N^{*}_{W}(C)\right)}{\partial C} = 0.$$

Thus the gradient of $MER_W(G_1,C)$ can be expressed in one of two ways:

$$\text{CASE1: } C < F^{c}_{S_{G0}}{}^{-1}\left(\frac{\pi_W}{\pi_G+p_G}\right) \Rightarrow \frac{\partial}{\partial C} MER_W(G_1,C) = (\pi_G+p_G)\cdot F^{c}_{S_{G0}}(C) \qquad \text{8B.7a}$$

$$\text{CASE 2: } C \geq F^{c}_{S_{G0}}{}^{-1}\left(\frac{\pi_W}{\pi_G+p_G}\right) \Rightarrow \frac{\partial}{\partial C} MER_W(G_1,C) =$$

$$(\pi_G+p_G)\int_{r_W=0}^{N^{*}_{W}(C)} F^{c}_{S_{G0}}(C-r_W) f_R(r_W)\,dr + \pi_W \cdot F^{c}_{r_W}\left(C-F^{c}_{S_{G0}}{}^{-1}\left(\frac{\pi_W}{\pi_G+p_G}\right)\right) \qquad \text{8B.7b}$$

Using the definition of $N_W^{*}(C)$ in Claim 8.1, it is easy to show that the value of expression 8B.7b is less than or equal to π_W. We can now substitute expression 8B.7 and the definition of $N_W^{*}(C)$ (equation 8.2) into 8B.3 to obtain two alternative expressions for the gradient of $ER_H(G_1,S_{H0},C,N_{H0})$:

$$\text{CASE 1: } N_{H0} < C-F^{c}_{S_{G0}}{}^{-1}\left(\frac{\pi_W}{\pi_G+p_G}\right):$$

$$\frac{\partial}{\partial N_{H0}}\left[\, ER_H(G_1,S_{H0},C,N_{H0})\right] =$$

$$\left(\pi_H + p_H - \pi_W F^{c}_{r_W}\left(C-N_{H0}-F^{c}_{S_{G0}}{}^{-1}\left(\frac{\pi_W}{\pi_G+p_G}\right)\right)\right) F^{c}_{S_{H0}}(N_{H0})$$

$$-(\pi_G+p_G)\int_{r_W=0}^{C-N_{H0}-F^{c}_{S_{G0}}{}^{-1}\left(\frac{\pi_W}{\pi_G+p_G}\right)} F^{c}_{S_{G0}}(C-N_{H_0}-r_W) f_R(r_W)\,dr \geq \pi_H + p_H - \pi_W \qquad \text{8B.8a}$$

CASE 2: $N_{H0} \geq C\text{-}F^{c}_{S_{G0}}{}^{-1}\left(\frac{\pi_W}{\pi_G+p_G}\right)$:

$$\frac{\partial}{\partial N_{H0}}\left[ER_H(G_1,S_{H0},C,N_{H0})\right] = \left(\pi_H+p_H - \left(\pi_G+p_G\right)F^{c}_{S_{G0}}(C-N_{H0})\right)F^{c}_{S_{H0}}(N_{H0}) \qquad 8B.8b$$

where the notation: $F^{c}_{X}(x) = 1-F_X(x)$ is used to represent the complementary cummulative distribution of random variable X at the point x.

In order to prove the claim, it remains to be shown that the set : $\Gamma = \left(N_{H0}: N_{H0} \geq 0 \text{ and } \nabla_{N_{H0}}[ER_H(G_1,S_{H0},C,N_{H0})] > 0\right)$ is convex. Since the complementary cummulative distribution functions of S_{H0}, S_{G0}, and r_W are non-decreasing, expressions 8.B.8a and 8B.8b are non-increasing in N_{H0}. If $C < F^{c}_{S_{G0}}{}^{-1}\left(\pi_W/\left(\pi_G+p_G\right)\right)$, then expression 8B.8a is irrelevant for $N_{H0} \geq 0$. If $C \geq F^{c}_{S_{G0}}{}^{-1}\left(\pi_W/\left(\pi_G+p_G\right)\right)$, then since $\pi_W < \pi_H+\pi_H$:

$$\nabla_{N_{H0}}[ER_H(G_1,S_{H0},C,N_{H0})] > 0, \text{ for } N_{H0} \in \left(0, C\text{-}F^{c}_{S_{G0}}{}^{-1}\left(\frac{\pi_W}{\pi_G+p_G}\right)\right)$$

At the point :

$$N_{H0} = C\text{-}F^{c}_{S_{G0}}{}^{-1}\left(\frac{\pi_W}{\pi_G+p_G}\right)$$

these expression 8B8.a is exactly equal to 8B8.b. It follows that the gradient of $ER_H(G_1,S_{H0},C,N_{H0})$ is non-increasing in N_{H0} , and the set $\Gamma = \left(N: N \geq 0 \text{ and } \nabla_N[ER_H(G_1,S_{H0},C,N)] > 0\right) = \left(0, N^{*}_{H0}\right)$ is convex.

QED

Chapter 9: Heuristics

As shown in the previous section, we have very good understanding of how to manage the reservations process on the target date. Once this date arrives, the number of outstanding guaranteed and 6 p.m. hold reservations has already been determined by the number and mix of requests that have been received and by management's willingness to accept them. Given these outstanding reservations, management must allocate rooms to 6 p.m. hold, walk-in, and guaranteed reservation customers as they arrive.

The first decision that must be made is the maximum number of rooms to allocate to 6 p.m. hold customers. Often it is much less expensive, both financially and in terms of good will, to walk a customer in the late afternoon than it would be later in the evening. Not only is it easier to locate an alternative hotel with availability, the customer is likely to experience much less discomfort. Thus, in certain situations when hotels are severely over-booked, they may elect to walk an early arrival in order to reduce the probability of having to walk some one else later on.

At 6 p.m., when the unclaimed hold reservations expire, the problem changes dramatically. Given that there are G_1 guaranteed reservations outstanding and that C_{WG} rooms have yet to be allocated, it is necessary to decide upon the number of of rooms to make available to walk-in customers. Note that by limiting the number of rooms that can be allocated to walk-ins, this decision implicitly protects rooms for the arrival of the outstanding guaranteed reservations.

Recall that we can solve this "walk-in" problem optimally. That is, given that after the arrival of the 6 p.m. holds there are G_1 guaranteed

reservations outstanding and there are C_{GW} rooms available, the optimal number of walk-ins to accept is as follows:

$$N_W^*(C_{WG}) = \begin{cases} 0 & \text{if } C_{WG} \le F_{S_{G0}}^{c\,-1}\left(\dfrac{\pi_W}{\pi_G + p_G}\right) \\ C_{WG} - F_{S_{G0}}^{c\,-1}\left(\dfrac{\pi_W}{\pi_G + p_G}\right) & \text{otherwise} \end{cases} \qquad 8.2$$

where $F_{S_{G0}}^{c\,-1}(\alpha) = G : \text{Prob}(S_{G0} \ge G) = \alpha$.

The intuition behind this result is simple. The marginal benefit from giving an additional room to walk-ins is the revenue π_W. The expected marginal cost of not having that room available for later arriving guaranteed reservations is equal to the probability = $\text{Prob}(S_{G0} > C_{WG} - N_W)$ that we will be unable to honor at least one reservation multiplied by the cost $(\pi_G + p_G)$ of doing so. Thus, it is in our interest to accept walk-ins as long as the probability of failing to honor a guaranteed reservation is less than $\pi_W/(\pi_G + p_G)$.

Under the mild restriction that the revenue from a walk-in be no greater than the lost revenue and penalty associated with failing to honor a 6 p.m. hold reservation, we can also solve the 6 p.m. room allocation problem to optimality. That is, given that the hotel has C rooms and there are G_1 outstanding guaranteed reservations, it is optimal to allocate rooms to the first N_{H0}^* customers to arrive with 6 p.m. hold reservations, where:

$$N^{*}_{H0} = \begin{cases} 0 & \text{if } \pi_H + p_H - (\pi_G + p_G) F^{c}_{S\infty}(C) < 0 \\ C - F^{c}_{S\infty}{}^{-1}\left(\dfrac{\pi_H + p_H}{\pi_G + p_G}\right) & \text{otherwise} \end{cases} \qquad 8.10$$

The intuition here is similar to that of the solution to the walk-in problem. In this case, we are trading off the certain increase in revenue and penalty savings of allocating a room to an arriving 6 pm hold customer against the expected cost of not having that room available later to allocate to a customer with a guarantee. This expected cost is equal to the revenue that is lost plus the penalty that is incurred by the failure to honor a guaranteed reservation weighted by the probability that at least one such failure will occur. With each additional room that is allocated to 6 p.m. holds, we increase the probability that the number of rooms remaining will not be sufficient to satisfy all of the guests who arrive with guarantees.

Unfortunately, it is not nearly so easy to obtain optimal solutions to the problem of placing limits on the numbers of guaranteed and 6 p.m. hold reservations to accept prior to the target date. The limits on the two types of reservations must be jointly determined. Since the two types may cancel and fail to show up at different rates, it is necessary to consider both the quantity and mix of reservations when deciding upon limits. Since the probability distributions for the number of customers who show up on the target date depends upon the number of reservations that have been taken, it is difficult to determine a closed form optimal solution to the reservations acceptance problem.

Because of the inherent difficulty of the reservation acceptance problem, managers rely on "rules of thumb" and "gut feel" to determine

when to begin refusing reservations requests. The heuristics which we describe were motivated by the insight which we obtained from analyzing the room allocation problem as well as from discussions that we held with managers from two large hotels near downtown Boston, The OMNI Parker House and The Marriott. During these discusions, it was our objective to understand the general approach and the "rules of thumb" that are used to solve the reservations acceptance problem in practice.

In each of the three heuristics, a target number of walk-ins is selected. This number represents the number of rooms that are targeted for walk-ins. It is based upon the size of the premium that a walk-in pays as well as the distribution function for walk-in requests. The target is as follows:

$$TargW = F_{rw}^{c\ -1}\left(\frac{\pi_R}{\pi_W}\right) \qquad 9.1$$

The intuition of this target is the marginal benefit associated with having an additional room for walk-ins is equal to the walk-in rate multiplied by the probability that demand for walk-ins would be sufficient to fill all of the rooms made available to them. As we make more rooms available to walk-ins (increasing TargW), the probability of filling them all ($F_{rw}^{c}(TargW)$) decreases. Since making an additional room available to walk-ins means forgoing the opportunity to fill it with a reservation customer, the opportunity cost of doing so is equal to the revenue associated with reservation customers. The marginal cost and benefit are balanced when the probability of having to turning at least one walk-in away is equal to the ratio of the two fares.

The first heuristic for setting limits on the numbers of guaranteed and 6 p.m. hold reservations to accept in period t is simple. Let us define

$ETrG_t$ to be the expected total number of requests for guaranteed reservations in periods t,...,1. Recall from the model of the problem that is described in chapter 6 that there are G_{t+1} and H_{t+1} reservations booked at the beginning of period t. The decisions N_{Gt} and N_{Ht} are the maximum numbers of guaranteed and hold reservations to accept during period t:

$H1_t(G_{t+1}, H_{T+1})$:

1. Calculate TargW using expression 9.1.
2. Set TargR = C - TargW.
3. Calculate $ETrG_t = E[r_{Gt} + ... + r_{G1}]$.

4a. If $G_{t+1} \cdot E[q_{G0}] + H_{t+1} \cdot E[q_{H0}] \geq$ TargR,

Then $N_{Gt} = N_{Ht} = 0$.

4b. Else:

$$N_{Gt} = \frac{\text{TargR} - G_{t+1} \cdot E[q_{G0}] - H_{t+1} \cdot E[q_{H0}]}{E[q_{G0}]},$$

$$N_{Ht} = \text{Max}\left\{0, \frac{\text{TargR} - (G_{t+1} + ETrG_t) \cdot E[q_{G0}] - H_{t+1} \cdot E[q_{H0}]}{E[q_{H0}]}\right\}$$

Steps 1 and 2 must be calculated only once. Steps 3 and 4 must be calculated in every period. The condition in step 4a is that, if there are no cancellations prior to the target date, the expected number of rooms that will be filled with reservation guests is at least as high as the target. Thus no reservations will be accepted in the next period. If the condition in 4a is not met, then guaranteed reservations are accepted until the point where the expected number of rooms that will be filled with reservation guests is equal to the target. 6 p.m. hold reservations are accepted only to the extent that the expected future demand ($ETrG_t$) for guaranteed reservations is

insufficient to meet the target. Notice that, with the exception of the targeted number of rooms for walk-ins, all calculations are made using expected values of the random variables. Also note that although heuristic H1 considers "no show" rates, it assumes that no one will cancel a reservation prior to the target date. Although this is consistent with the "rules of thumb" that we observed in practice, we propose a modified heuristic, H2, which explicitly considers the effects of attrition, i.e. early cancellations, upon reservation limits. Let us define $EYrG_t$ to be the expected total "yield" from requests for guaranteed reservations in periods t,...,1, where yield refers to the number of requests net of cancellations and no-shows. Also, let q_{Gt} (q_{Ht}) be the probability that a customer with a guaranteed (6 p.m. hold) reservation does not cancel during period t. The rule is as follows:

$H2_t(G_{t+1}, H_{t+1})$:

1. Calculate TargW using expression 9.1.
2. Set TargR = C - TargW.
3. Calculate $EYrG_t = E\left[\sum_{\tau=1}^{t}\left(r_{G\tau}\prod_{i=0}^{\tau-1} q_{Gi}\right)\right]$

4a. If $G_{t+1}\cdot\prod_{\tau=0}^{t} q_{G\tau} + H_{t+1}\cdot\prod_{\tau=0}^{t} q_{H\tau} \geq TargR$,

Then $N_{Gt} = N_{Ht} = 0$.

4b. Else:

$$N_{Gt} = \frac{TargR - G_{t+1} \cdot \prod_{\tau=0}^{t} q_{G\tau} - H_{t+1} \cdot \prod_{\tau=0}^{t} q_{H\tau}}{\prod_{\tau=0}^{t-1} q_{G\tau}},$$

$$N_{Ht} = Max\left\{0, \frac{TargR - G_{t+1} \cdot \prod_{\tau=0}^{t} q_{G\tau} - EYrG_t - H_{t+1} \cdot \prod_{\tau=0}^{t} q_{H\tau}}{\prod_{\tau=0}^{t-1} q_{H\tau}}\right\}.$$

H2 is identical to H1 except for the fact that it explicitly inflates the reservations limits to allow for cancellations prior to the target date. The third heuristic which we consider differs from H2 only in the way in which TargR , the number of rooms that are targeted to be filled with reservation customers, is calculated.

$H3_t(G_{t+1}, H_{t+1})$:

1. Calculate TargW using expression 9.1.
2. Set ε such that: $Prob\left[\frac{C - TargW}{E[q_{G0}]} \cdot q_{G0} \le C - TargW + \varepsilon\right] = \frac{p_G}{\pi_G + p_G}$

   Set TargR = C - TargW - ε.

Steps 3 and 4 as in heuristic H2.

The intuition behind step 2 of heuristic H3 is similar to that of the standard newsboy problem. The number of rooms targeted for reservations is deflated until the probability of being unable to honor a reservation is equal to the ratio of the cost of overbooking versus the sum of the costs of overbooking plus the lost revenue from an empty room. In step 2 of heuristic H3, ε represents the number of additional rooms that would be required on the target date in order to provide a service level of $p_G / (\pi_G +$

p_G) if exactly (C - TargW) / $E[q_{G0}]$ guaranteed reservations and TargW walk-ins were accepted . Thus, after reducing by ε the number of rooms targeted for reservations, the probability of being unable to honor a reservation reflects the trade-off between the cost of doing so and that of an empty room.

In order to evaluate the three heuristics that we have described, we have used Monte Carlo simulation to estimate the expected cost of using each one. For each reservation acceptance heuristic, we assume that, on the target date, the optimal room allocation policies (expressions 8.2 and 8.10) are followed. In the following section, we describe a lower bound on the expected cost of the original dynamic program which models the hotel reservations problem. This bound serves as a basis for comparison with the simulation of the heuristics. In section 11, we describe the parameters of the simulation and discuss the results.

Chapter 10: An Upper Bound for the Dynamic Program

In order to evaluate the performance of the heuristics, it is necessary to have an upper bound on the value of an optimal solution to the dynamic program $MER_t(G_{t+1},H_{t+1},C)$. We present such a bound and show how it can be obtained by successively bounding the functions: $ER_G(G_1,C_G)$, $MER_W(G_1,C_{WG})$, $MER_0(G_1,H_1,C),...,MER_T(G_{T+1},H_{T+1},C)$.

Before, introducing the upper bound, we will show that the function $R_G(S_{G0},C_G)$ can be expressed as the optimal value of a linear program. This result will be used in the subsequent proof.

Claim 10.1: The value of the function $R_W(G_1,r_W,C_{WG},N_W)$ =

$$\pi_W W + ER_G(G_1,C_{WG}-W) \qquad 7.7a$$

$$\text{st:} \quad W = \text{Min}(N_W, r_W) \qquad 7.7b$$

is less than or equal to the optimal value of the following mathematical program:

$UR_W(G_1,r_W,C_{WG},N_W)$ =

$$\text{Max} \quad \pi_W W + ER_G(G_1,C_{WG}-W) \qquad 10.1a$$

$$\text{s.t.} \quad W \le r_W \qquad 10.1b$$

$$W \le N_W \qquad 10.1c$$

$$W \ge 0 \qquad 10.1d$$

where the decision variable W represents the actual number of walk-ins to arrive and receive rooms given that the limit is N_W and there are r_W requests.

Proof: Let $W = \text{Min}(r_W, N_W)$. Then W is a feasible solution to the maximization problem in $UR_W(G_1,r_W,C_{WG},N_W)$. Suppose that W* is an optimal solution to this problem. Then:

$$UR_W(G_1,r_W,C_{WG},N_w) = \pi_W W^* + ER_G(G_1,C_{WG}-W^*) \geq$$

$$\pi_W W + ER_G(G_1,C_{WG}-W) = R_W(G_1,r_W,C_{WG},N_w) \qquad \text{QED}$$

Using this preliminary result, we can derive an upper bound on the value of an optimal solution to $MER_t(G_{t+1},H_{t+1},C)$ by succesively developing upper bounds for the later stages of this dynamic program.

Claim 10.2: The solution to the following linear program is an upper bound on the function $MER_W(G_1,C_{WG})$. That is:

$$MER_W(G_1,C_{WG}) \leq UMER_W(G_1,C_{WG}) =$$

$$E_{r_W,S_{G0}}\Big[Z_W(r_W,S_{G0},C_{WG}) =$$

$$\begin{aligned}
&\text{Max} \quad \pi_W N_W + \pi_G N_{GA} - p_G N_{GW} && 10.2a\\
&\text{s.t.: } N_W \leq r_W && 10.2b\\
&\qquad N_W \leq C_{WG} && 10.2c\\
&\qquad N_W + N_{GA} - C_{WG} \leq 0 && 10.2d\\
&\qquad N_{GA} + N_{GW} - S_{G0} = 0 && 10.2e\\
&\qquad N_W, N_{GA}, N_{GW} \geq 0 \ \Big] && 10.2f
\end{aligned}$$

Proof: Using claim 10.1, we have,

$$MER_W(G_1,C_{WG}) = \underset{N_W \leq C_{WG}}{\text{Max}} \{ER_W(G_1,C_{WG},N_W)\}$$

$$\leq \underset{N_W \leq C_{WG}}{\text{Max}} \{E_{r_W}[UR_W(G_1,r_W,C_{WG},N_W)]\} \qquad 10.3$$

We make the following observation:

$$\underset{N_W \leq C_{WG}}{\text{Max}} \{E_{r_W}[UR_W(G_1,r_W,C_{WG},N_W)]\}$$

$$\leq E_{r_W}\Big[\underset{N_W \leq C_{WG}}{\text{Max}} \{UR_W(G_1,r_W,C_{WG},N_W)\}\Big] \qquad 10.4$$

The intuition behind the inequality above is that, in the expression on the left hand side, the expectation is taken with respect to a single maximizing value of N_W. In the expression on the right hand side, the maximization with respect to N_W is performed for each possible realization of the random variable r_W. Similarly, we observe that:

$$UR_W(G_1, r_W, C_{WG}, N_W) = \underset{W \in \Omega_W}{\text{Max}} \left\{ \pi_W W + E_{S_{G0}}[R_G(S_{G0}, C_{WG}-W)] \right\}$$

$$\leq E_{S_{G0}} \left[\underset{W \in \Omega_W}{\text{Max}} \left\{ \pi_W W + R_G(S_{G0}, C_{WG}-W) \right\} \right] \qquad 10.5$$

where Ω_W represents the intersection of constraints 10.1b-d. Substituting this result into expression 10.4, we have:

$$MER_W(G_1, C_{WG}) = \underset{N_W \leq C_{WG}}{\text{Max}} \{ER_W(G_1, C_{WG}, N_W)\}$$

$$\leq E_{r_W} \left[\underset{N_W \leq C_{WG}}{\text{Max}} \{UR_W(G_1, r_W, C_{WG}, N_W)\} \right]$$

$$\leq E_{r_W} \left[\underset{N_W \leq C_{WG}}{\text{Max}} \left\{ E_{S_{G0}} \left[\underset{W \in \Omega_W}{\text{Max}} \left\{ \pi_W W + R_G(S_{G0}, C_{WG}-W) \right\} \right] \right\} \right] \qquad 10.6$$

Once again, we can interchange the maximization and expectation functions without changing the direction of the inequality. In particular:

$$MER_W(G_1, C_{WG}) \leq E_{r_W} \left[\underset{N_W \leq C_{WG}}{\text{Max}} \left\{ E_{S_{G0}} \left[\underset{W \in \Omega_W}{\text{Max}} \left\{ \pi_W W + R_G(S_{G0}, C_{WG}-W) \right\} \right] \right\} \right]$$

$$\leq E_{r_W,S_{G0}}\left[\max_{\substack{N_W \leq C_{WG} \\ W \in \Omega_W}} \left\{\pi_W W + R_G(S_{G0}, C_{WG}-W)\right\}\right] \qquad 10.7$$

The maximization that is inside of the expectation in expression 10.7 is equivalent to the following linear program:

$Z_W(r_W,S_{G0},C_{WG}) =$

Max:	$\pi_W W + \pi_G N_{GA} - p_G N_{GW}$	10.8a
s.t.:	$N_W \leq C_{WG}$	10.8b
	$W \leq r_W$	10.8c
	$W \leq N_W$	10.8d
	$W + N_{GA} - C_{WG} \leq 0$	10.8e
	$N_{GA} + N_{GW} - S_{G0} = 0$	10.8f
	$N_W, W, N_{GA}, N_{GW} \geq 0$	10.8g

Finally, we observe that for any optimal solution to 10.8a-g, we can set $N_W = W$ without affecting either its optimality or feasibility. Thus, after eliminating the unnecessary constraint 10.8d, and substituting N_W for W everywhere, it can be seen that the above linear program is equivalent to the one in the claim. QED

Before proposing an upper bound for the function $MER_0(G_1,H_1,C)$, we prove the following claim:

Claim 10.3: For $N_{H0} \leq C$, the function:

	$R_H(G_1,S_{H0},C,N_{H0}) = \pi_H N_{HA} - p_H N_{HW} + MER_W(G_1,C-N_{HA})$	7.5a
s.t.:	$N_{HA} = \text{Min}(N_{H0}, S_{H0})$	7.5b
	$N_{HA} + N_{HW} = S_{H0}$	7.5c

is less than or equal to the optimal value of the following mathematical program:

$UR_H(G_1,S_{H0}(H_1),C,N_{H0})$ =

Max: $\pi_H N_{HA} - p_H N_{HW} + MER_W(G_1,C-N_{HA})$ 10.9a

s.t.: $N_{HA} \leq N_{H0}\ (\leq C)$ 10.9b

$N_{HA} + N_{HW} = S_{H0}$ 10.9c

$N_{HA}\ ,\ N_{HW} \geq 0$ 10.9d

where N_{HA} is a random variable representing the number of customers with 6 pm hold reservations who are assigned rooms. It is dependent upon S_{H0} , the number of 6 pm hold customers to arrive, and upon N_{H0}, the limit on the number that will be allocated rooms. N_{HW} represents the number whose reservations are not honored. Note that, in the definition of $UR_H(G_1,S_{H0}(H_1),C,N_{H0})$, we have shown H_1 as a parameter of the random variable S_{G0} to emphasize the dependency between the number of customers to arrive with 6 p.m. hold reservations and the number of outstanding reservations. To simplify the notation, we will omit this parameter in future references to the random variable S_{H0}.

Proof: Let $N_{HA} = Min(N_{H0}, S_{H0})$, and let $N_{HW} = Max\ (0, S_{H0} - N_{HA})$ for some realization of the random variable S_{H0}. This is equivalent to the conditions 7.5b and 7.5c. This solution is feasible in 10.9b-d, and the value of the objective function (10.9a) at this point is equal to $R_H(G_1,S_{H0},C,N_{H0})$. Thus, the the value of an optimal solution to 10.9a-d is at least this large.

QED

We are now ready to propose an upper bound on $MER_0(G_1,H_1,C)$:

Claim 10.4: The solution to the following linear program is an upper bound on the function $MER_0(G_1,H_1,C)$:

$UMER_0(G_1,H_1,C)$ =

$$E_{S_{H0},rw,S_{G0}}\Big[\ \text{Max}\ \pi_H N_{HA} - p_H N_{HW} + \pi_W N_W + \pi_G N_{GA} - p_G N_{GW} \qquad 10.10a$$

s.t.:

$$N_W \le r_W \qquad 10.10b$$

$$N_{HA} + N_W \le C \qquad 10.10c$$

$$N_{HA} + N_W + N_{GA} \le C \qquad 10.10d$$

$$N_{GA} + N_{GW} = S_{G0} \qquad 10.10e$$

$$N_{HA} \le N_{H0}\ (\le C) \qquad 10.10f$$

$$N_{HA} + N_{HW} = S_{H0} \qquad 10.10g$$

$$N_{H0} \le C \qquad 10.10h$$

$$N_{HA}, N_{HW}, N_W, N_{GA}, N_{GW} \ge 0\ \Big] \qquad 10.10i$$

Proof: Recall that $MER_0(G_1,H_1,C)$ represents the maximization of the function $ER_H(G_1,S_{H0},C,N_{H0})$ with respect to N_{H0}, where $ER_H(G_1,S_{H0},C,N_{H0})$ is the expectation of the function $R_H(G_1,S_{H0},C,N_{H0})$ with respect to S_{H0}. Using claim 10.3:

$$MER_0(G_1,H_1,C) = \underset{\substack{N_{H0}\le C \\ N_{H0}\ge 0}}{\text{Max}} \{ER_H(G_1,S_{H0},C,N_{H0})\}$$

$$= \underset{\substack{N_{H0}\le C \\ N_{H0}\ge 0}}{\text{Max}} \{E_{S_{H0}}[R_H(G_1,S_{H0},C,N_{H0})]\}$$

$$\le \underset{\substack{N_{H0}\le C \\ N_{H0}\ge 0}}{\text{Max}} \{E_{S_{H0}}[UR_H(G_1,S_{H0},C,N_{H0})]\} \qquad 10.11$$

Interchanging the maximization and expectation functions does not change the direction of the inequality:

$$MER_0(G_1,H_1,C) \le E_{S_{H0}}\left[\underset{\substack{N_{H0}\le C \\ N_{H0}\ge 0}}{\text{Max}} \{UR_H(G_1,S_{H0},C,N_{H0})\}\right] \qquad 10.12$$

After replacing $UR_H(G_1,S_{H0},C,N_{H0})$ with its mathematical programming formulation (equations 10.9a-d), we have:

$$MER_0(G_1,H_1,C) \leq$$

$$E_{S_{H0}}\left[\underset{\substack{N_{H0}\leq C\\ N_{H0}\geq 0}}{\text{Max}}\left\{\underset{N_{HA},N_{HW}\in\Omega_{RH}}{\text{Max}}\left(\pi_H N_{HA} - p_H N_{HW} + MER_W(G_1,C-N_{HA})\right)\right\}\right]$$

$$\leq E_{S_{H0}}\left[\underset{\substack{N_{H0}\leq C\\ N_{H0}\geq 0\\ N_{HA},N_{HW}\in\Omega_{RH}}}{\text{Max}}\left(\pi_H N_{HA} - p_H N_{HW} + UMER_W(G_1,C-N_{HA})\right)\right]$$

$$= E_{S_{H0}}\left[\underset{\substack{N_{H0}\leq C\\ N_{H0}\geq 0\\ N_{HA},N_{HW}\in\Omega_{RH}}}{\text{Max}}\left(\pi_H N_{HA} - p_H N_{HW} + E_{r_W,S_{G0}}[Z_W(r_W,S_{G0},C-N_{HA})]\right)\right] \qquad 10.13$$

where Ω_{RH} represents the intersection of constraints 10.9b-d, and $Z_W(r_W,S_{G0},C-N_{HA})$ represents the optimal solution to the linear program given in equations 10.2 a-f. Once again, we can take the expectation with respect to r_W and S_{G0} outside of the maximization without altering the direction of the inequalities: After expanding $Z_W(r_W,S_{G0},C-N_{HA})$ to its linear programming formulation, we have the result we have sought:

$$MER_0(G_1,H_1,C) \leq UMER_0(G_1,H_1,C) =$$

$$E_{S_{H0},r_W,S_{G0}}\Big[\text{Max } \pi_H N_{HA} - p_H N_{HW} + \pi_W N_W + \pi_G N_{GA} - p_G N_{GW} \qquad 10.10a$$

s.t.: constraints 10.10 b - i. $\Big]$

QED.

With the preceding results, we now propose an upper bound on the value of an optimal solution to the dynamic program $MER_T(G_{T+1},H_{T+1},C)$. Recall that vectors are indicated by bold face notation, e.g.: $\mathbf{r}_\tau = (r_{G\tau},r_{H\tau})$ for $\tau = 1,...,T$, and $\mathbf{S}_\tau = (S_{G\tau},S_{H\tau})$ for $\tau = 0,...,T$.

Claim 10.5: Let $\widehat{H}_t = r_{Ht} + S_{Ht}(\widehat{H}_{t+1})$ and $\widehat{G}_t = r_{Gt} + S_{Gt}(\widehat{G}_{t+1})$ for $t = 1,...,T$. An optimal solution to $MER_T(G_{T+1},H_{T+1},C)$ is less than or equal to $UMER_T(G_{T+1},H_{T+1},C)$, the solution to the following problem:

$$E_{r_T,...,r_1,r_W,S_T,...,S_0}\Big[\ \text{Max}\ \pi_H N_{HA} - p_H N_{HW} + \pi_W N_W + \pi_G N_{GA} - p_G N_{GW} \qquad 10.14a$$

s.t.:

$$N_{HA} + N_{HW} \le S_{H0}(\widehat{H}_1) \qquad 10.14b$$

$$N_W \le r_W \qquad 10.14c$$

$$N_{GA} + N_{GW} \le S_{G0}(\widehat{G}_1) \qquad 10.14d$$

$$N_{GA} + N_{HA} + N_W \le C \qquad 10.14e$$

$$N_{HA}, N_{HW}, N_{GA}, N_{GW}, N_W \ge 0\ \Big] \qquad 10.14f$$

Note that the expectation is taken with respect to r_W and the vectors $\mathbf{r}_t$ for $t = T,...,1$ and $\mathbf{S}_t$ for $t = T,...,0$.

Proof: Recall the definition of problem $MER_T(G_{T+1},H_{T+1},C)$. If we perform the maximization with respect to N_{GT} and N_{HT} inside of the expectation, the result is at least as large as $MER_T(G_{T+1},H_{T+1},C)$:

$MER_T(G_{T+1},H_{T+1},C) =$

$$\underset{N_{GT},N_{HT}}{\text{Max}}\{ER_T(G_{T+1},H_{T+1},C,N_{GT},N_{HT})\} =$$

$$\underset{N_{GT},N_{HT}}{\text{Max}}\left\{\sum_{r_{GT}=0}^{\infty}\sum_{r_{HT}=0}^{\infty}\sum_{S_{GT}=0}^{G_{T+1}}\sum_{S_{HT}=0}^{H_{T+1}} MER_{T-1}(G_T,H_T,C)f_{r_T,S_T}(\mathbf{r}_T,\mathbf{S}_T)\right\}$$

$$\leq \sum_{r_{GT}=0}^{\infty}\sum_{r_{HT}=0}^{\infty}\sum_{S_{GT}=0}^{G_{T+1}}\sum_{S_{HT}=0}^{H_{T+1}}\left\{\underset{N_{GT},N_{HT}}{\text{Max}}\{MER_{T-1}(G_T,H_T,C)f_{r_T,S_T}(\mathbf{r}_T,\mathbf{S}_T)\}\right\} \qquad 10.15$$

The intuition of the inequality in 10.15 is that, if we knew in advance the number of requests and cancellations that would be received during the forthcoming period, we could make at least as good a decision about the number of requests to accept as we could in the absence of this knowledge. By recursively applying this result to $MER_{T-1}(G_T,H_T,C)$,..., $MER_1(G_2,H_2,C)$, we have:

$MER_T(G_{T+1},H_{T+1},C) \leq$

$$\sum_{\mathbf{r}_T,..,\mathbf{r}_1}\sum_{\mathbf{S}_T,..,\mathbf{S}_1}\left\{\underset{\substack{N_{GT},...,N_{G1}\\N_{HT},...,N_{H1}}}{\text{Max}}\{MER_0(G_1,H_1,C)f_{r,S}(\mathbf{r}_T,...,\mathbf{r}_1,\mathbf{S}_T,..,\mathbf{S}_1)\}\right\}$$

$$= E_{\substack{r_T,..,r_1\\s_T,..,s_1}}\left[\underset{\substack{N_{GT},...,N_{G1}\\N_{HT},...,N_{H1}}}{\text{Max}}\{MER_0(G_1,H_1,C)\}\right] \qquad 10.16$$

where the bold face indicates vectors: $\mathbf{r}_t = (r_{Gt}, r_{Ht})$ and $\mathbf{S}_t = (S_{Gt}, S_{Ht})$, and the reservation inventories follow the recursive relationship:

$$G_t = \text{Min}\{r_{Gt}, N_{Gt}\} + S_{Gt}(G_{t+1}), \text{ for } t = 1,...,T$$

$$H_t = \text{Min}\{r_{Ht}, N_{Ht}\} + S_{Ht}(H_{t+1}), \text{ for } t = 1,...,T$$

Note that we have written G_{t+1} (H_{t+1}) as a parameter of S_{Gt} (S_{Ht}) to emphasize the dependency between the number of reservations to survive until point t and the number in inventory at t+1. After substituting the upper bound on the value of $MER_0(G_1,H_1,C)$:

$$MER_T(G_{T+1},H_{T+1},C) \le E_{\substack{r_T,\ldots,r_1 \\ s_T,\ldots,s_1}}\left[\max_{\substack{N_{GT},\ldots,N_{G1} \\ N_{HT},\ldots,N_{H1}}}\{UMER_0(G_1,H_1,C)\}\right] \qquad 10.17$$

Recall the definition of $UMER_0(G_1,H_1,C)$ in equations 10.10a-i. It is the expected value of a linear program that is parameterized by the random variables S_{H0}, S_{G0}, and r_W. As we have done throughout, we can take the expectation with respect to these variables outside of the maximization over N_{GT}, N_{HT},..., N_{G1}, N_{H1} without affecting the direction of the inequality in equation 10.17. Thus, after substituting for $UMER_0(G_1,H_1,C)$ we have:

$$MER_T(G_{T+1},H_{T+1},C) \le$$

$$E_{\substack{r_T,\ldots,r_1,r_W \\ s_T,\ldots,s_0}}\left[\max_{\substack{N_{GT},\ldots,N_{G1} \\ N_{HT},\ldots,N_{H1}}}\left\{\begin{array}{l}\max\ \pi_H N_{HA} - p_H N_{HW} + \pi_W N_W + \pi_G N_{GA} - p_G N_{GW} \\ \text{s.t. constraints 10.11b-h}\end{array}\right\}\right]$$

$$= E_{\substack{r_T,\ldots,r_1,r_W \\ s_T,\ldots,s_0}}\Big[\max\ \pi_H N_{HA} - p_H N_{HW} + \pi_W N_W + \pi_G N_{GA} - p_G N_{GW} \qquad 10.18a$$

s.t.:

$$G_t = \min\{r_{G,t}, N_{G,t}\} + S_{G,t},\ \text{for } t = 1,\ldots,T \qquad 10.18b$$

$$H_t = \min\{r_{H,t}, N_{H,t}\} + S_{H,t},\ \text{for } t = 1,\ldots,T \qquad 10.18c$$

$$N_{GT},\ldots,N_{G1},N_{HT},\ldots,N_{H1} \ge 0 \qquad 10.18d$$

$$N_W \le r_W \qquad 10.10b$$

$$N_{HA} + N_W \le C \qquad 10.10c$$

$$N_{HA} + N_W + N_{GA} \le C \qquad 10.10d$$

$$N_{GA} + N_{GW} = S_{G0} \qquad 10.10e$$

$$N_{HA} \le N_{H0}\ (\le C) \qquad 10.10f$$

$$N_{HA} + N_{HW} = S_{H0} \qquad 10.10g$$

$$N_{H0} \le C \qquad 10.10h$$

$$N_{HA},N_{HW},N_W,N_{GA},N_{GW} \ge 0 \Big] \qquad 10.10i$$

This is the expected value of the optimal solution to a linear program that is parameterized by the random variables r_{Gt} and r_{Ht} (for t = T,...,1), S_{Gt}, S_{Ht} (for t = T,...,0) and r_W. It remains to be shown that this value is less than or equal to the upper bound that was proposed in equations 10.14a-f. We propose to do so by showing that the expression in the claim represents the expected value of a less tightly constrained linear program than the one above.

It is easy to show that, regardless of the values of decision variables N_{Ht} and N_{Gt} (t = 1,...,T), the inventories of 6 p.m. holds and guaranteed reservations will satisfy the following conditions in period 1:

$$H_t = \text{Min}\{r_{Ht}, N_{Ht}\} + S_{Ht} \le r_{Ht} + S_{Ht}, \quad \text{for } t = 1,...,T \qquad 10.19a$$

$$G_t = \text{Min}\{r_{Gt}, N_{Gt}\} + S_{Gt} \le r_{Gt} + S_{Gt}, \quad \text{for } t = 1,...,T \qquad 10.19b$$

Thus, regardless of the values of the decision variables (N_{HT}, ...,N_{H1} and N_{GT}, ...,N_{G1}), for any given realization of the random variables representing random requests for walk-ins and reservations, cancellations, and no-shows, we have that: $S_{G0}(\widehat{G}_1) \ge S_{G0}(G_1)$ and $S_{G0}(\widehat{H}_1) \ge S_{G0}(H_1)$. It follows that expression 10.14b is less restrictive than the combination of 10.18c and 10.10h. Similarly, 10.14c is less restrictive than the combination of 10.18b and 10.10e.

After substituting these looser constraints, we have the following relaxation of the above linear program:

$$\text{Max } \pi_H N_{HA} - p_H N_{HW} + \pi_W N_W + \pi_G N_{GA} - p_G N_{GW} \qquad 10.18a$$

s.t.:

$$N_{HA} + N_{HW} \le S_{H0}(\widehat{H}_1) \qquad 10.14b$$

$$N_{GA} + N_{GW} \le S_{G0}(\widehat{G}_1) \qquad 10.14d$$

$$N_W \le r_W \qquad 10.11b$$

$$N_{HA} + N_W \le C \qquad 10.11c$$

$$N_{HA} + N_W + N_{GA} \le C \qquad 10.11d$$

$$N_{HA} \le N_{H0} \ (\le C) \qquad 10.11f$$

$$N_{H0} \le C \qquad 10.11h$$

$$N_{HA}, N_{HW}, N_W, N_{GA}, N_{GW} \ge 0 \qquad 10.11i$$

$$N_{G,T-1}, \ldots, N_{G1}, N_{H,T-1}, \ldots, N_{H1} \ge 0 \qquad 10.20$$

Note that the decision variables $N_{GT},\ldots,N_{G1},N_{HT},\ldots,N_{H1}$, and N_{H0} are now irrelevant, and constraint 10.20 can be ignored. After eliminating the redundant constraints (10.10c, 10.10f, and 10.10h), it is easy to see that the above linear program is equivalent to the one in equations 10.14a-f.

QED

Each of the upper bounds on the various stages of the dynamic program are based upon interchanging maximization with expectation. When the maximization is performed inside of the expectation function, it is possible to maximize with respect to each different realization of the random variables. In other words, if a hotel manager had the benefit of hindsight and could make all of his reservation acceptance and room allocation decisions after observing all of the requests, cancellations, and no-shows, he could make better decisions than he can in real life. The bound which is described in this section represents the expected value of an optimal solution to the problem if it were solved with the aid of hindsight.

This bound provides a benchmark against which we can evaluate the performance of heuristic methods of solving the real problem where decisions must be made before the random events (reservations requests, cancellations, and no-shows) have been observed. In the following section we describe the parameters of a Monte Carlo Simulation of the heuristics, and discuss their performance.

Chapter 11: Computational Results

In order to evaluate the performance of the reservation acceptance heuristics that are described in Chapter 8, we used Monte Carlo simulation. We simulated the performance of each heuristic in an environment in which reservations demand, cancelations, and no-shows and the number of walkins were random. It was assumed that, on the target date, the optimal room allocation policies were followed. By measuring the hotel profits over a series of repeated simulations of this environment, we were able to obtain statistical estimates of the expected costs of using each heuristic.

We also used Monte Carlo simulation to determine a statistical estimate of the upper bound. Recall that the upper bound is the expected optimal value of a linear program in which the coefficients in the right hand side of the constraint matrix are random. We "estimated" this expected optimal value by repeatedly generating realizations of the random coefficients and solving the resulting linear programs. This "estimate" of the upper bound on the expected revenue of an optimal solution to the original dynamic program provides a benchmark against which we can evaluate the performance of the heuristics.

In the Monte Carlo simulations, we assumed that in each decision period prior to the target date, the numbers of requests for 6 p.m. holds and guaranteed reservations are drawn from Poisson distributions. On the target date, the number of requests for walk-ins was also assumed to be a Poisson distributed random variable. We assumed that both no-shows and cancellations of reservations are drawn from Binomial distributions. Given the assumption that each customer acts independently, these

distributions are intuitively appealing. Their use is also common in the literature. For example, Alstrup et. al. claim that the sales of airline tickets follow Poisson distributions while cancellations and no-shows follow Binomials. Rothstein (1974) defends the validity of these distributions for modeling hotel reservations processes, and uses them to test his results.

We developed our tests on the basis of discussions with the management of a Marriott Hotel near downtown Boston. The parameters that we used approximate those at this hotel. The target dates which present the management of this hotel with the most difficulty are the ones for which there is considerable uncertainty regarding demand, i.e. Friday, Saturday, and Sunday. Because of the uncertainty, a large number of discounts are made available, and the average price paid by a customer with either a 6 p.m. hold or a guaranteed reservation is about $100 versus $150 for a walk-in. The cost of failing to honor a reservation before 6 p.m. is $100 versus $250 later in the evening. Although a large component of these costs represents intangibles, the sustantial difference between them can be explained by the fact that prior to 6.pm., it is often possible to re-locate a customer to another downtown Marriott. By attempting to re-locate customers prior to the point when there are no more rooms available, the hotel affords itself the luxury of offering several customers the option of re-locating. It is not uncommon to be able find someone who is quite receptive to spending a free night at an alternative hotel. Alternatively, when the re-location occurs later in the evening, and customers are given no option, it is usually perceived as a far greater inconvenience.

For the purposes of the simulations, we assumed that there are four decsion points prior to the target date, and that the length of the intervals between decisions is chosen such that that the expected number of

reservation requests is the same in each of the four periods. For our base case scenario, we assumed that the probability distributions have the parameters given in Table 11.1:

t	$E[r_G]$	$E[r_H]$	$E[r_W]$	q_G	q_H
4	70	40	-----	.9	.9
3	70	40	-----	.9	.9
2	70	40	-----	.9	.9
1	70	40	-----	.9	.9
0	-----	-----	30	.9	.5

Table 11.1: Scenario A

The first column in Table 11.1 is the number of periods that remain before the target date. The second, third, and fourth columns contain the expected number of requests for guaranteed reservations, 6 p.m. holds, and walk-ins in each period. Note that no requests for reservations are made on the target date, and that requests for walk-ins occur only then. Columns five and six contain the probability, q_{Gt} (q_{Ht}) that a given guaranteed (6 p.m. hold) reservation that is held at the beginning of period t will not cancel before the end of period t. Recall that on the target date, period 0, a cancellation is equivalent to a no-show.

The other scenarios that we tested were variations on this base case. The parameters of these scenarios are given in Appendix 11A. In scenario B, the expected number of requests for guaranteed reservations in each period leading up to the target date was 100 instead of 70. In scenario C, this parameter was 40. In scenario D we returned the expected demand for

guaranteed reservations to the base case level, but increased the expected number of requests for walk-ins to 60.

For each of these scenarios, we ran 500 iterations of the heuristics. By taking sample averages over the 500 iterations, we obtained statistical estimates of the expected costs of using the heuristics in each of the four different conditions. These estimates are presented in Table 11.2 as percentages of the upper bound. The standard deviations associated with these estimates is roughly .1%.

Recall that the calculation of the upper bound is itself a statistical estimate of the expected value of a linear program in which the right hand side coefficients are random variables. This estimate is also based upon 500 generations of the random variables, and has a standard deviation of roughly .1%.

Scenario	H1	H2	H3	Upper Bound
A	94.5%	97.3%	97.3%	$31,262
B	93.3%	97.1%	97.3%	$31,496
C	99.9%	99.9%	99.9%	$23,646
D	94.7%	96.8%	96.8%	$32,978

Table 11.2: Revenue of Heuristics as a % of the Upper Bound

It can be seen in Table 11.2 that H2 and H3 perform much better than H1 for scenarios A, B, and D. Since H1 does not consider the effects of reservation attrition, i.e. cancellations prior to the target date, it tends to accept fewer reservations in the planning periods furthest from the target date than do the other heuristics. As should be expected, all three heuristics perform very well for scenario C. In this scenario the expected demand for rooms,

net of cancellations and no-shows, is very low, and nearly everyone who wants a room receives one. Thus, hindsight is of little benefit, and the performance of the heuristics is very close to that of the upper bound.

Using expected revenue as a measure of performance, heuristic H3 offers only marginally better performance than H2. In an attempt to further differentiate these heursics, we considered another measure of performance: the frequency with which at least five reservations cannot be honored. Although the hotel managers with whom we spoke indicated that their primary concern is maximizing revenue, they are also concerned about being unable to honor large numbers of reservations because of overbooking. If we consider the likelihood that a large number of reservations cannot be honored, then H3 looks more attractive. Table 11.3 contains the number of target dates out of 500 in which at least five customer reservations could not be honored. Although by this measure, H1 dominates the other heuristics, the expected revenue from using it indicates that it might be too conservative. However, H3 has expected profits that are comparable to those of H2, and simultaneously results in fewer occasions in which large number of reservations cannot be honored.

Scenario	H1	H2	H3
A	8	44	42
B	1	61	44
C	0	0	0
D	13	49	43

Table 11.3: Occasions (out of 500) in which at least 5 reservations could not be honored.

To test the robustness of the heuristics with respect to the probability that a customer with guaranteed reservation will fail to show-up on the target date ($1 - q_{G0}$), we performed a series of tests in which this probability varied between .05 and .3. Because we wanted to focus our attention on the effect of increased variablity rather than decreased total demand for rooms, we also adjusted the parameters of the requests for reservations. In particular, we adjusted the parameters for reservations requests so that the product of the expected number of requests and probability that in each test:

$$\sum_{t=1}^{T} \left(E[r_{Gt}] \cdot \prod_{\tau=0}^{t-1} q_{G\tau} \right) = K_G$$

where K_G is a constant (≈ 216.657). For example, for $q_{G0} = .75$ instead of .90, the expected number of requests for guaranteed reservations in each period prior to the target date was 84 instead of 70. The results of these simulations are presented in Table 11.4 and 11.5:

$1 - q_{G0}$	H1	H2	H3	Upper Bound
.05	94.5%	97.8%	97.8%	$31,250
.1 (Base Case)	94.5%	97.3%	97.3%	$31,262
.15	94.5%	97.0%	97.1%	$31,289
.2	94.7%	96.8%	96.9%	$31,237
.25	94.9%	96.5%	96.7%	$31,201
.3	94.7%	96.5%	96.7%	$30,246

Table 11.4: Revenue of Heuristics as a % of the Upper Bound

It is interesting to observe in Table 11.4 that the heuristics performance relative to the upper bound does not seem to be adversely affected when the prabability that a guaranteed reservation will show-up decreases.

However, in table 11.5, it can be seen that, for H2 and H3, the frequency with which at least 5 customers must be re-located increases

q_{G0}	H1	H2	H3
.95	2	26	24
.9 (Base Case)	8	44	42
.85	2	60	56
.8	6	67	56
.75	5	88	73
.7	3	79	63

Table 11.5: Occasions (out of 500) in which at least 5 reservations could not be honored.

with the no-show probability (1 - q_{G0}) for a guaranteed customer. However, this frequency does not seem to increase as quickly for H3 as for H2. For a guaranteed reservation no-show probability of 1 - q_{G0} = .05, H3 results in 8% fewer occasions with at least five customer re-locations than H2. For 1 - q_{G0} = .3, H3 results in 20% fewer occasions.Note that, regardless of the no-show probability for guaranteed reservations, H1 results in very few occasions where large numbers of customer reservations cannot be honored. However, it achieves this at the expense of expected revenues by turning away large numbers of reservations.

Although heuristics H1, H2, and H3 all perform well with respect to the upper bound under a variety of of conditions, H2 and H3 perform consistently better than H1. This advantage results from the fact that H2 and H3 consider the effects of cancellations prior to the target date, while H1 considers only no-shows in determining the number of reservations to accept. H3 is slightly more sophisticated than H2 in that it adds a safety

factor into the number of rooms targeted to be filled with reservations customers. This safety factor is based on the trade-off between the lost revenue of an empty room and the penalty for overbooking. Although the performance of H3 was only slightly better than that of H2 in terms of expected revenue, it resulted in substantially fewer occasions in which large numbers of customers had to be re-located because of overbooking. By this latter measure, its advantage over H2 increased as the show-up prabability for the guaranteed reservations decreased.

Appendix 11A

t	$E[r_G]$	$E[r_H]$	$E[r_W]$	q_G	q_H
4	100	40	-----	.9	.9
3	100	40	-----	.9	.9
2	100	40	-----	.9	.9
1	100	40	-----	.9	.9
0	-----	-----	30	.9	.5

Table 11.6: Scenario B

t	$E[r_G]$	$E[r_H]$	$E[r_W]$	q_G	q_H
4	40	40	-----	.9	.9
3	40	40	-----	.9	.9
2	40	40	-----	.9	.9
1	40	40	-----	.9	.9
0	-----	-----	30	.9	.5

Table 11.7: Scenario C

t	$E[r_G]$	$E[r_H]$	$E[r_W]$	q_G	q_H
4	70	40	-----	.9	.9
3	70	40	-----	.9	.9
2	70	40	-----	.9	.9
1	70	40	-----	.9	.9
0	-----	-----	60	.9	.5

Table 11.8: Scenario D

Chapter 12: Discussion

The problems that arise in the hotel industry as a result of the uncertainties associated with room reservations are both interesting and complex. Although there are obvious similarities between these "yield management" problems and the ones faced by the airline industry, there is also an interesting parallel with the problems that arise from random yields in manufacturing.

In manufacturing, there is often uncertainty as to the fraction of a production batch that will conform to the customer's specifications. Lot sizes should be determined so as to minimize the expected costs of backorders (or lost sales) and of holding inventory. In some cases, different customers require different levels of performance, and the individual units in a given production batch can be classified and sold as different products on the basis of performance. In such cases, known as co-production, it may also be possible to downgrade, i.e. to satisfy a customer with a product that meets a higher specification than he requires. Thus management must make two types of decisions: lotsizing and product allocation decisions.

In the hotel industry, management attempts to maximize the revenues associated with a future target date by accepting and refusing requests for an assortment of types of reservations. Based on the prices and probabilities of cancellations or no-shows that are associated with the various types of reservations, it may be desirable to restrict the sales of certain types. The problem of determining limits on various types of reservations is similar to the one of determining lot sizes in manufacturing. In both environments, a lot size (reservation limit) is selected in order to balance the expected costs of inventory and backorders

(empty rooms and overbooking penalties). However, because the number of reservations depends upon demand, a hotel has much less control over reservations than a manufacturer has over a lot size.

On the target date of the reservations, rooms must be allocated to customers as they arrive. Because there are differences in the prices and overbooking penalties associated with different types of reservations, this problem is similar to the one faced by managers of co-production processes when they can down-grade products. The analogy is even closer when hotels have more than one type of room.

One of the objectives of this thesis is to bring some of the expertise which has been developed in the manufacturing environment to bear upon the hotel reservations problem. Based on interaction with the managements of two hotels near downtown Boston, the OMNI Parker House and the Marriott, we focused our efforts on one dimension of the overall problem that has not, to our knowledge, been studied previously. The particular problem which we have studied arises from the fact that rooms must be allocated to customers as they arrive throughout the target date. In other words, management must begin allocating rooms before observing the number of no-shows. Note that this contrasts sharply to the situation in the airline industry where all of the no-shows are observed prior to the boarding of the airplane.

To facilitate our study of this dimension of the problem, we have identified three distinct types of customers, those with 6 p.m. hold reservations, those with credit card guarantees, and walk-ins. We assume that there is only one type of room and that reservations are for single night stays. The model that we propose is a stochastic dynamic program in which the stages are the time periods prior to a target date, and the states

are the numbers of booked guaranteed and 6 p.m. hold reservations. The final two stages of the dynamic program represent the allocation of rooms to the customers as they arrive.

After providing optimal solutions to the final stages of the dynamic program, i.e. the room allocation problems, we suggest heuristic methods of solving the reservations acceptance problems in the earlier stages. The heuristics are based upon both mathematical analysis and the interaction we had with the managements of two large urban hotels. In order to obtain a benchmark against which to evaluate the heuristics, we derive an upper bound on the value of an optimal solution to the dynamic program. Finally, we evaluate the performance of the heuristics using Monte Carlo simulation.

Although all three of the heuristics for reservations acceptances perform well relative to the upper bound, two of them (H2 and H3) stand out. All three heuristics set a target number of rooms to be filled with customers with reservations as opposed to walk-ins. H1 limits the acceptance of reservations based on this target and the probability that a reservation will fail to show-up. However, the better heuristics (H2 and H3) also factor in the probability that reservations will cancel prior to the target date. As a result, they they tend to accept more reservations than does H1.

The performances of H2 and H3 were comparable in terms of expected revenue. Even in the worst cases that we tested, both heuristics resulted in expected revenues that were about 97% of the upper bound. However, H3 resulted in fewer occasions in which there were at least five re-locations, i.e. customers who had to be re-located to another hotel because their reservations could not be honored. For a no-show probability of .05, H2 resulted in at least five re-locations on 5.12% of the simulated

target days, compared to 4.8% for H3. This advantage became more pronounced when the probability of no-shows increased. For a no-show probability of .3, H2 resulted in at least five re-locations 15.8% the simulated target days, versus 12.6% for H3. The difference between the performance of these two rules can be attributed to the fact that H3 reduces its reservation limits by a "safety margin" that is based on the trade-off between the lost revenue from an empty room against the penalty for having to re-locate a customer whose reservation cannot be honored.

These results represent an encouraging first step in the study of an interesting dimension of the hotel reservations problem which has not, to our knowledge, been studied previously. The most obvious direction for future research would be to extend these results to cases that are not covered by our two most restrictive assumptions, i.e. one type of room, and only single night stays. We conjecture that the room allocation problem could be solved for multiple room types using a decomposition approach. For example, we might consider the capacity of each type of room as a separate "mini-hotel". When the allocation problem is solved for room type i, reservations for other room types would be treated similarly to the way that walk-ins are treated in the single room type model. Clearly, such an extension would add a great deal of complexity to the model. As such, it should be incorporated into the model only if doing so facilitates insight into the underlying real-world problem.

It has been the objective of this thesis to develop practical approaches to problems that arise as a result of uncertainty in two surprisingly similar environments: semi-conductor manufacturing, and hotel reservations. In each case, the problems can be modeled as stochastic dynamic programs. By analyzing the mathematics of these models and interacting with

practitioners, we develop heuristic methods of solving the problems. Then, using the intuitive concept that an omniscient decision maker can do at least as well as one who must make decisions prior to observing random outcomes, we develop bounds on the optimal values of these programs. Using these bounds as benchmarks, we show via Monte Carlo simulation that our heuristics perform well.

Although a great deal of quantitative research has been performed with respect to manufacturing industries, there has been surprisingly little done in the services. Thus, our original intent was to bring manufacturing expertise to bear upon a problem in the service industry. We were surprised to discover that the benefits went in both directions. By simultaneously studying two related problems, that arise in very different environments, we were able to gain a deeper understanding of each one.

REFERENCES

Allen, F. M., R. Braswell, and P. Rao, "Distribution-free Approximations for Chance-Constraints," Operations Research, 22, pp. 610-621, 1974.

Alstrup, Jens, Soren Boas, Oli B.G. Madsen, and Rene Victor Valqui, 1986. "Booking Policy for Flights with Two Types of Passengers," European Journal of Operations Research, Vol. 27, pp. 274-288.

Avram, Florian and Lawrence Wein, "A Product Design Problem in Semi-Conductor Manufacturing," Working Paper, MIT Sloan School of Management, 1990.

Belobaba, P.P. 1987. "Airline Yield Management - An Overview of Seat Inventory Control," Transportation Science, Vol. 21, pp. 63-73.

Belobaba, P.P. 1989. "Application of a Probabiliistic Decision Model to Airline Seat Inventory Control," OR Practice, Vol. 37, No. 2, pp. 183-197.

Baker, Kenneth R., Michael J. Magazine, and Henry L. W. Nuttle, "The Effect of Commonality on Safety Stock in a Simple Inventory Model," Management Science, Vol. 32, No. 8, pp.982-988, 1986.

Birge, John R., "The Value of the Stochastic Solution in Stochastic Linear Programs with Fixed Recourse," Mathematical Programming, 24, pp. 314-325, 1982.

Bitran, G. R., and S. Dasu, "Ordering Policies in an Environment of Stochastic Yields and Substitutable Demands," M.I.T. Slaon School Working Paper #3019-89-MS, 1989.

Bitran, G. R., and T.Y. Leong, "Deterministic Approximations to Co-Production Problems with Service Constraints," M.I.T. Slaon School Working Paper #3071-89-MS, 1990.

Bitran, G. R., and D. Tirupati, "Planning and Scheduling for Epitaxial Wafer Production Facilities," Operations Research, 36, pp.34-49, 1988.

Deuermeyer, Bryan L. and William P. Pierskalla, "A By-Product Production System with an Alternative," Management Science, Vol. 24, No. 13, pp.1373-1383, 1978.

Drake, Alvin W., Fundamentals of Applied Probability Theory, McGraw-Hill Book Company, 1967.

Gerchak, Y., R. G. Vickson, and M. Parlar, "Periodic Review Production Models with Variable Yield and Uncertain Demand," IIE Transactions, (20) 2, 1988.

Kothari, V., "Silicon Wafer Manufacture," Unpublished thesis, M.I.T. Sloan School of Management, May, 1984.

Ladany, Shaul 1976. "Dynamic Operating Rules for Motel Reservations," Decision Sciences, Vol. 7, pp. 829-840.

Lee, Hau L., "Lot Sizing in a Production Process with Defective Items, Process Corrections, and Rework." Working Paper, Stanford University, Department of Industrial Engineering and Engineering Management, August, 1990.

Liberman, Varda and Uri Yechiali 1978. "On the Hotel Overbooking Problem - An Inventory System with Stochastic Cancellations," Management Science, Vol. 24, No. 11, pp. 1117-1126.

Mangasarian, Olvi L. Nonlinear Programming, McGraw-Hill Book Company, 1969.

Mazzola, Joseph B., William F. McCoy, and Harvey M. Wagner "Algorithms and Heuristics for Variable-Yield Lot Sizing," Naval Research Logistics, Vol. 34, pp.67-88, 1987.

Ou, Jihong, and Lawrence M. Wein, "Dynamic Scheduling of a Production/Inventory System with Bi-Products and Random Yield," M.I.T. Operations Research Center Technical Report #250-91, May, 1991.

Rothstein, Marvin 1985. "OR and the Airline Overbooking Problem," OR Forum, Vol. 33, No. 2, pp. 237-248.

Sanchez, M. and R. Martinez, "Automatic Booking Level Control," Procedings of the Tenth Annual Symposium of the Airline Group of the International Federation of Operational Research Societies, 1970, published by American Airlines Inc., New York

Singh, Medini, Chacko Abraham, and Ramakrishna Akella, "A Wafer Design Problem in Semi-Conductor Manufacturing for Reliable Customer Service", IEEE Transactions on Components, Hybrids, and Manufacturing Technology. Vol 13, No 1, March 1990.

Symonds, G.H., "Deterministic Solutions for a Class of Chance Constrained Programming Problems," Operations Research, 15, pp. 495-512, 1967.

Tang, Christopher S., "The Impact of Uncertainty on a Production Line", Management Science, Vol. 36, No 12, December, 1990.

Williams, Fred E. 1977. "Decsion Theory and the Innkeeper: An Approach for Setting Hotel Reservation Policy," Interfaces, Vol. 7, No. 4, pp. 18-31.

Yano, Candace and Hau L. Lee, "Lot Sizing with Random Yields; A Review," Working Paper, The University of Michigan, Department of Industrial and Operations Engineering, 1989.