OPTIMAL INVENTORY MODELING OF SYSTEMS

Multi-Echelon Techniques

Second Edition

Recent titles in the

INTERNATIONAL SERIES IN
OPERATIONS RESEARCH & MANAGEMENT SCIENCE
Frederick S. Hillier, Series Editor, *Stanford University*

Ramík, J. & Vlach, M. / *GENERALIZED CONCAVITY IN FUZZY OPTIMIZATION AND DECISION ANALYSIS*
Song, J. & Yao, D. / *SUPPLY CHAIN STRUCTURES: Coordination, Information and Optimization*
Kozan, E. & Ohuchi, A. / *OPERATIONS RESEARCH/ MANAGEMENT SCIENCE AT WORK*
Bouyssou et al. / *AIDING DECISIONS WITH MULTIPLE CRITERIA: Essays in Honor of Bernard Roy*
Cox, Louis Anthony, Jr. / *RISK ANALYSIS: Foundations, Models and Methods*
Dror, M., L'Ecuyer, P. & Szidarovszky, F. / *MODELING UNCERTAINTY: An Examination of Stochastic Theory, Methods, and Applications*
Dokuchaev, N. / *DYNAMIC PORTFOLIO STRATEGIES: Quantitative Methods and Empirical Rules for Incomplete Information*
Sarker, R., Mohammadian, M. & Yao, X. / *EVOLUTIONARY OPTIMIZATION*
Demeulemeester, R. & Herroelen, W. / *PROJECT SCHEDULING: A Research Handbook*
Gazis, D.C. / *TRAFFIC THEORY*
Zhu, J. / *QUANTITATIVE MODELS FOR PERFORMANCE EVALUATION AND BENCHMARKING*
Ehrgott, M. & Gandibleux, X. /*MULTIPLE CRITERIA OPTIMIZATION: State of the Art Annotated Bibliographical Surveys*
Bienstock, D. / *Potential Function Methods for Approx. Solving Linear Programming Problems*
Matsatsinis, N.F. & Siskos, Y. / *INTELLIGENT SUPPORT SYSTEMS FOR MARKETING DECISIONS*
Alpern, S. & Gal, S. / *THE THEORY OF SEARCH GAMES AND RENDEZVOUS*
Hall, R.W./*HANDBOOK OF TRANSPORTATION SCIENCE - 2nd Ed.*
Glover, F. & Kochenberger, G.A./*HANDBOOK OF METAHEURISTICS*
Graves, S.B. & Ringuest, J.L. / *MODELS AND METHODS FOR PROJECT SELECTION: Concepts from Management Science, Finance and Information Technology*
Hassin, R. & Haviv, M./ *TO QUEUE OR NOT TO QUEUE: Equilibrium Behavior in Queueing Systems*
Gershwin, S.B. et al/ *ANALYSIS & MODELING OF MANUFACTURING SYSTEMS*
Maros, I./ *COMPUTATIONAL TECHNIQUES OF THE SIMPLEX METHOD*
Harrison, T., Lee, H. & Neale, J./ *THE PRACTICE OF SUPPLY CHAIN MANAGEMENT: Where Theory And Application Converge*
Shanthikumar, J.G., Yao, D. & Zijm, W.H./*STOCHASTIC MODELING AND OPTIMIZATION OF MANUFACTURING SYSTEMS AND SUPPLY CHAINS*
Nabrzyski, J., Schopf, J.M., Węglarz, J./ *GRID RESOURCE MANAGEMENT: State of the Art and Future Trends*
Thissen, W.A.H. & Herder, P.M./ *CRITICAL INFRASTRUCTURES: State of the Art in Research and Application*
Carlsson, C., Fedrizzi, M., & Fullér, R./ *FUZZY LOGIC IN MANAGEMENT*
Soyer, R., Mazzuchi, T.A., & Singpurwalla, N.D./ *MATHEMATICAL RELIABILITY: An Expository Perspective*
Talluri, K. & van Ryzin, G./ *THE THEORY AND PRACTICE OF REVENUE MANAGEMENT*
Kavadias, S. & Loch, C.H./*PROJECT SELECTION UNDER UNCERTAINTY: Dynamically Allocating Resources to Maximize Value*
Sainfort, F., Brandeau, M.L., Pierskalla, W.P./ *HANDBOOK OF OPERATIONS RESEARCH AND HEALTH CARE: Methods and Applications*
Cooper, W.W., Seiford, L.M., Zhu, J./ *HANDBOOK OF DATA ENVELOPMENT ANALYSIS: Models and Methods*

**** A list of the early publications in the series is at the end of the book ****

OPTIMAL INVENTORY MODELING OF SYSTEMS

Multi-Echelon Techniques

Second Edition

by
Craig C. Sherbrooke, Ph.D.

KLUWER ACADEMIC PUBLISHERS
NEW YORK, BOSTON, DORDRECHT, LONDON, MOSCOW

eBook ISBN: 1-4020-7865-X
Print ISBN: 1-4020-7849-8

Created in the United States of America

Dedication

This book is dedicated to
Rosalie, the next generation of
mathematicians Andrew and
Evan, and the following
generation Joshua and Michael

Contents

List of Figures

List of Tables

List of Variables

The variables below will sometimes carry subscripts as defined in the text. We have used three letter mnemonics for probability density functions, abbreviated pdf below, except that p is used for the Poisson. Random variables are abbreviated r.v. The symbol ^ indicates an estimated value.

ROMAN LETTERS

a	Parameter of a pdf or type of event
A	Availability
b	Parameter of a pdf or type of event
BO	r.v. for backorders
bin(x)	Binomial pdf of x
c	Item cost ($)
C	Cost of system spares ($)
d	Demand/quarter
D	Sum of daily demand rates at bases/number bases
DI	r.v. for stock due-in
e	2.718..(Euler's constant)
E[X]	Expected value of the random variable X
EB	Estimated backorders from regression (Appendix B)
EBO(s)	Expected backorders with a stock level s
EFR(s)	Expected fill rate with a stock level s
erl(x)	Erlang pdf of x
exp(t)	Exponential pdf of t
$Ex(x)$	Expected number of periods with x demands
f	Fraction or probability
F	Flying hours/quarter
$g(x)$	Probability of x aircraft (end-items) grounded/down

$G(x)$	Cumulative probability of x or fewer aircraft down
gam(t)	Gamma pdf of t
$h(x)$	Probability of x
$H(x)$	Cumulative probability of x or less
hyp(x)	Hypergeometric pdf of x
i	Index for item number
I	Total number of items
j	Index for base number or system number
J	Total number of bases (sites)
k	Protection level, degrees of freedom for the chi-square
K	Number of systems ($\leq N$) that must operate
L	Inventory position (on-hand + due-in – backorders)
lap(x)	Laplace pdf of x
log	Natural logarithm (base e)
LB	Lower bound on backorders (Appendix B)
m	Average annual demand
n	Number of time periods, number of trials
N	Total number of systems or aircraft (end-items)
neg(x)	Negative binomial pdf of x
NRTS	Not Repairable This Site (1- r)
O	Average order and ship time
Obs(x)	Number of periods with x demands observed
OH	r.v. for stock on-hand
$p(x)$	Poisson pdf of x
P(x)	Cumulative Poisson pdf of x or less
$\Pr\{X = x\}$	Probability that r.v. X equals the value x
q	Probability that failure of an item is due to this child
Q	Order quantity
r	Probability of base repair
R	Reorder point
s	Stock level
$S(K)$	Probability that systems 1, 2 . . K operate; K + 1, . . N down
t	Time
T	Average repair time
u	Units on hand
UB	Upper bound on backorders (Appendix B)
v	Volume
V	Variance/mean ratio of demand
Var[X]	Variance of the random variable X
VBO(s)	Variance in backorders with a stock level s
w	Weight
W	Set of probabilities defined in Equation 6.6

wei(x)	Weibull pdf of x
x	Number of demands, number in pipeline
X	Random variable for number of demands, number in pipeline
y	Number of demands, number in pipeline
Y	Random variable for number of demands, number in pipeline
z	Minimum number of locations of an item for parent operation
z	Total number of locations for an item in its parent item, $z \leq Z$

LOWER-CASE GREEK LETTERS

α	[a] , the integer $\leq$ a
β	Backorder target for an item
δ	Demands/flying hour
χ^2	Chi square probability distribution
ι	Interest rate
κ	Shrinking constant
λ	Lagrange multiplier
μ	Average demand over the lead time, average pipeline
π	Annual cost of a backorder ($)
$\varphi(x)$	Cumulative probability of x or fewer backorders due to LRUs or SRUs
ψ	Equation 5.34
ρ	Probability of demand, correlation of demand
σ	Standard deviation (square root of the variance)
τ	Time
ω	Exponential smoothing constant

UPPER CASE GREEK LETTERS

$\Delta[h(x)]$	First difference $h(x+1) - h(x)$
$\Gamma(x)$	The gamma function, defined as $x!$ for integral x.
Ω	Order cost ($)

Preface

This book is written for the logistician who is concerned with one or more systems or end equipments and with the percent of time that they are operational. We develop the mathematical modeling techniques to determine the optimal inventory levels by item and location for any specified system availability or total spares investment. The optimizations consider trade-offs between stock at the operating locations and the supporting depots, known as *the multi-echelon problem*; between stock for an item and its sub-items, known as *the multi-indenture problem.*

In addition, this book is written for the graduate student in operations research who is interested in the mathematics of inventory theory and its application to real problems. The theoretical foundations of the requisite inventory theory are covered in detail. As the sub-title indicates, multi-echelon (and multi-indenture) techniques are an important part of the book. We believe this is the first text to consider these topics in depth.

However, this is not primarily a book on multi-echelon inventory theory. We restrict our attention in the optimization theory to the case where the stock level is s and a reorder or repair of one unit is initiated whenever the level falls to $s - 1$. This is the only policy that we consider, because it is the optimal policy for the high-cost, low-demand repairable items of which systems are composed. We do calculate order quantities that can be larger than one for low cost, high demand items. However, because these items appear at lower indentures in the parts hierarchy, we are content to use approximations to the optimal policy, knowing that the impact on system availability and system cost will be slight.

The reader who is primarily interested in the mathematics of the general multi-echelon problem should refer to other sources. The classic reference is Clark and Scarf (1960), and there is an excellent anthology by Schwarz (1981). Several of the papers in the Schwarz anthology deal with the multi-echelon problem, including over 200 references. Other more recent works of note include Federgruen and Zipkin (1984) and Svoronos and Zipkin (1988). Due to the complex iterative nature of the solution techniques for these optimal, multi-echelon policies, there have been few applications to date. An important exception, Cohen et al (1990), is discussed in Chapter 10.

In the past twenty years there have been two important, conflicting developments in the management of inventories. The manufacturing sector has tended to place more emphasis on better planning and "just-in-time" methods to reduce investment in in-process inventories. At the other extreme, logisticians who are responsible for the support of complex equipments such as ships, telecommunications networks, electric utilities, computer systems, space shuttles and orbiting vehicles are making use of ever more sophisticated inventory models. This is due in part to the increasing complexity of these equipments, and the need to meet specified availability targets. Central to both developments has been the tremendous increase in computing power, computer literacy and widespread user access.

Demand forecasting and inventory modeling are becoming less important to the former group, while they are becoming more critical to the latter. Between the extremes there are many other applications, such as those for retailers in the commercial world. In some cases retailers have been able to shorten lead times, and depend on greater responsiveness from their suppliers; in others, the variability of lead times and the number of wholesale suppliers has been increasing. Inventory theory and forecasting may still be important for them, but there is less of a need for new and better techniques.

Our objective in this book is to address the problem of supporting high-technology equipments. Though many of the most natural applications are in the military sector, the techniques that we develop are appropriate for complex civilian programs, too. Rather than talk in abstract terms about high-technology equipments and retail sites, it will be convenient in our discussions to adopt military examples and refer to aircraft, operating bases, supporting depots, etc. We hope this will make the context clearer and less academic without causing the reader to ignore other applications.

The stimulus for writing this book was a four-day (now three day) course on spares management and modeling that I first presented in April 1989.

The course has now been presented over fifty times in various locations in the United States, Europe, and the Far East. Before each subsequent course, the material was revised to reflect students' comments and the author's experience. The current form owes much to the feedback from hundreds of students.

The attendees have ranged from logisticians and engineers with extensive experience and doctoral degrees to managers with limited mathematical training. The book is intended to appeal to a similar audience with a range of interests and ability. Many of the mathematical proofs are placed in the Problems and Appendices to make the text easier for the reader who has less mathematical facility. (Calculus is unavoidable in a few sections of Chapters 4, 5, and 7, however).

I have taught inventory theory courses in graduate schools of operations research - usually using *Analysis of Inventory Systems* by Hadley and Whitin (1963) as the principal text. That book contains some excellent material, though it is out of date and out of print. However, students complained that Hadley and Whitin and other texts had few real examples, and they wanted to know more about whether the models had been implemented. Consequently, this book includes actual data from field tests of the techniques, demand prediction studies, and from work for Space Station *Freedom* wherever possible. Furthermore, every model discussed in the book has been programmed on personal computers, and most are being used today.

It is important to emphasize that the models developed in this book are all analytic. Simulation is used to verify the accuracy of the analytic models, but the models themselves consist of mathematical equations that can be solved for optimal stockage policies in an efficient manner. The analytic nature of the models is essential for practical application on personal computers or even mainframes.

The book includes a careful development of the mathematical foundations of the theory, appropriate for a one-semester graduate course. It is the author's hope that the book will be used both by practicing logisticians who want to keep up with the state of the art in inventory modeling, and by graduate students of operations research who are interested both in theory and practice.

A large part of the material in the book is based on my research. Much of it has been published in journals such as *Operations Research, Management Science,* and *the Naval Research Logistics Quarterly.* However, the modeling of periodic resupply for Space Station *Freedom,* where there is redundancy at both the system and item levels, is too recent to have appeared in print. Much of the demand prediction work has been described only in Logistics Management Institute publications.

The material in this book begins with research performed in the early 1960s. The research showed that it was possible to operate an Air Force base and achieve higher performance at significantly less cost for spares. Subsequently the research findings were validated in a field test in which the recommended stocks were actually implemented at the base, resulting in the same performance at about half the inventory cost. The philosophical basis for this new approach is given in Chapter 1, and the mathematical techniques in Chapter 2. It is shown that minimizing the sum of base backorders is equivalent to maximizing availability.

In Chapter 3 the mathematics is extended to the joint optimization of stock levels at bases and at the supporting depot. Chapter 4 treats demand rate estimation, and suggests techniques to model demand rates that do not stay constant. We show that this results in larger variance-to-mean ratios than the value of 1 that characterizes the Poisson distribution. The negative binomial distribution is used to model this effect as well as the larger variance-to-mean ratios that occur because the pipeline delays between echelons and indentures are not independent. This is illustrated with a two-indenture example. We describe demand prediction studies using actual Air Force data, and present methods for dealing with items whose failures are dominated by wear out. In Chapter 5 we develop the mathematics for the combination multi-echelon, multi-indenture optimization problem.

Chapter 6 and 7 are concerned with the periodic resupply problem for repairable items, and its application to Space Station *Freedom*. One of the new results presented in this book for the first time is an optimization technique where redundancy is modeled at both the system and item level. This has important implications in the design of systems. The same model can be used for long-term procurement problems and for short-term resupply manifesting of the space shuttle; in the latter case the age of installed units subject to wear out can be used to improve the set of items resupplied in a given shuttle flight.

In some applications, maintenance performs cannibalization: consolidation of item shortages on the smallest number of end-items. The mathematics for cannibalization is different; this is the subject of Chapter 8. We show that it is possible to use the same objective function, expected availability, though the results are only quasi-optimal. We note that regardless of the procurement model used, it is possible to achieve better short-term performance if information on the location and condition of assets at each point in time is used in decision-making. The DRIVE (Distribution and Repair in Variable Environments) model for distribution of serviceable assets from depot to bases and for prioritization of repair at depot is such a model. We describe some of the benefits and problems of implementing such a technique. Chapter 9 is new in the second edition, describing a dozen problems that can be modeled with the same theory, including modifications for commercial airlines, variations in the resupply and repair assumptions, treating sites that operate aircraft and support other sites.

Finally, Chapter 10 is concerned with many of the real-world problems of using models. What are the advantages and what are the limitations? Implementation experiences by several different user groups are presented. The appendices provide mathematical proofs of Palm's theorem, and discussions of special topics such as lateral supply between bases and demand prediction studies. Appendices D-F are new in the second edition. Appendix D is concerned with predicting spares demand in a wartime environment, based on observations from Desert Storm. Appendices E and F describe implementations of the optimization theory (VMetric) and the demand prediction theory (Demand Analysis System).

This book differs from other books on inventory theory in several important ways. We use the system approach, whereby we focus on the availability of the end-items such as aircraft, and then determine the appropriate inventory policies. We believe that logisticians should provide management with cost-availability curves, from which an optimal system target can be chosen. In fact, the system approach is used in several ways - not just in the determination of stockage levels but in demand prediction and in the evaluation of alternative policies.

Repairable items are the central focus here, because they most directly relate to aircraft availability, whereas consumable items are the focus in most books. We devote a lot of time to multi-echelon, multi-indenture inventory theory, though these are only given a couple of summary pages in most texts

Although only four chapters and appendices are totally new in this second edition, I have made extensive revisions in all chapters, adding numerous worked-out examples. The first edition was published twelve years ago, and many things have changed since that time as reflected in the new edition. For example, the personal computer models in 1992 did not use Windows, now the standard; the original book was done in WordStar, not Word, requiring an archaelogical project on the part of my son, Evan, to reconstruct the original manuscript.

I can be reached by email at csherbro@alum.mit.edu.

CRAIG SHERBROOKE

Camarillo, *CA.*

Acknowledgements

The materials in this book that were developed by me were performed under the sponsorship of several organizations. As a graduate student at the Massachusetts Institute of Technology during 1958-1960, I worked on Army inventory problems. From 1962-1969 at the RAND Corporation, much of the material in Chapters 1-4 was developed for the Air Force. From 1981-1993 I was a consultant to the Logistics Management Institute where the material in Chapters 5 and 8 as well as Appendices B-D was done for the Air Force; Chapters 6 and 7 for NASA and Space Station *Freedom.* At other times the author has worked on Navy and Defense Logistics Agency studies, and consulted with private companies on inventory problems.

It is impossible to thank everyone who has influenced and helped me, because of the large number of such people. My earliest productive work was largely spurred by a collaboration with George Feeney in my first days at the RAND Corporation, under the supportive management of the late Murray Geisler. Later at the Logistics Management Institute (LMI) I was fortunate to work with Mike Slay, one of the most creative logistics modelers I have known. Rob Kline worked with me on the Space Station application, and T. J. O'Malley supervised the research for the Air Force and NASA. The DRIVE model was developed jointly with Jack Abell and Lou Miller of RAND. Several others deserve thanks for encouraging me to write the book including Saul Gass, Jack Muckstadt, Ben Blanchard, and Rod Stewart. Bob Butler should be included in this category, because he first urged me to teach the course on which the book is based. My mathematician sons, Andrew and Evan, suggested several changes to the notation and exposition, all of which were incorporated. The notation would have been far more confusing but for the patience of the editor, Isabel Stein, among whose many contributions was the insistence that a given symbol have the same meaning from one chapter to the next.

Finally I want to thank LMI and its former President, Bob Pursley, for providing me with some time to write the first edition; thanks also go to several colleagues who critiqued individual chapters including Chris Hanks, Rob Kline, and Sal Culosi. Mike Slay spent many hours patiently looking for errors and suggesting improvements throughout the book. Though I am responsible for all remaining errors of commission and omission, I am most grateful that so many have were eliminated by their efforts.

The second edition was motivated by the large number of things that have happened in spares logistics over the past eleven years. I have added nearly a hundred pages including a new chapter, Chapter 9, three new appendices (D-F), and substantial revision and expansion of several chapters including more worked-out examples. Unfortunately, the first edition had a large number of typos and some substantive errors in Chapter 6 concerning finite populations. I apologize because it is hard enough to read an advanced text without encountering errors.

The new edition would not have happened without the strong support of Fred Hillier, whose distinguished career in operations research is well known. I appreciate the help of many people in updating the book including Rich Moore, Bob McCormick, Norm Scurria, Jim Russell, Meyer Kotkin, Sal Culosi, Randy King, and Mike Slay who brought me up-to-date on implementation by the services; to Ken Woodward, the architect of the VMetric interface and much more, who assisted in getting the latest information on VMetric; and to my wife Rosalie who has become a wizard at downloading pictures. Deborah Doherty and others at Kluwer helped me to overcome the sometimes mystifying ways of Microsoft Word and the Kluwer templates where objects can appear and disappear capriciously.

C.C.S.

Chapter 1

INTRODUCTION

Finally we shall place the Sun himself at the center of the Universe. All this is suggested by the systematic procession of events and the harmony of the whole Universe, if only we face the facts as they say "with both eyes open".

–Nicolas Copernicus

1.1 Chapter Overview

We introduce the fundamental notion of the system approach and contrast it with the older, traditional method of calculating spares known as the item approach. We show that for high technology equipments, repairable items are more important than consumable or non-repairable items.[1] This makes for some simplification, because there is only one decision variable on each item: when to order or, equivalently, the stock level. On the other hand, the problem is more complicated, because the support of complex systems

[1] We use the term *repairable* to signify items that have some possibility of being repaired. The military services use the term *reparable* to mean an item that may be repairable, depending on the nature of the failure. Nonmilitary readers are apt to think that the military term is a misspelling, so we prefer not to use it or the word *recoverable*, which means the same thing.

requires us to be concerned about many items and stock levels at both bases and depots. This is known as a *multi-echelon* context. Furthermore, we want to optimize the mix of items and the sub-items of which they are composed, known as the *multi-indenture* problem.

The terms "failure" and "demand" are used interchangeably. We assume that when there is a demand a spare is needed. If no spare is on hand, some system has a "hole" and the "end item" is unavailable until a spare can be supplied. Instead of using the term end item we will use aircraft as a typical example, and a military context where these models first arose. But, it is important to realize that the models we develop in this text have many commercial applications including commercial airlines, power plants, radar installations, space station. In fact, the theory is applicable to any complex system where it is meaningful to talk about *availability* (the percent of time that the system is operational). The system of interest may not be the aircraft, say, but a sub-system of the aircraft such as the guidance, the propulsion, or the avionics. We use the term "item" to designate a specific type of part and "units" for the quantity of the item. We will show that the stock level on any item at any location can be thought of as the average number of units of the item in repair or resupply plus some safety level to protect against variability in the demand and repair processes. But the optimal stock level depends on a number of other variables also including item cost, location (base or depot), and indenture (item or sub-item) as well as system variables such as the desired availability. In later chapters we develop the theory necessary to include all of these factors.

We summarize field test experience using a variable protection level that demonstrated as much as fifty percent reduction in inventory cost to obtain the same performance level, even at a single base. Last, the chapter shows with a simple example what optimal item policies look like. We show that the optimal stock levels are different when there is cannibalization - consolidation of aircraft "holes" or backorders to the smallest number of aircraft possible by remove-and-replace maintenance

1.2 The System Approach

In *the system approach,* questions are asked such as: How can we insure that 95% of our scheduled aircraft flights will not be delayed for lack of spare parts? How much more money do we need to spend to move from 95% to something higher? More generally what can we do to change our logistics support structure to achieve a desired availability more efficiently? Is it economic to have more repair capability at the operating sites?

The perspective of a retailer such as Sears Roebuck is very different. Retailers are interested in measures of customer satisfaction such as *fill rate*, the fraction of demands that are met from stock on the shelf. If the customer demand can not be met, there are two possibilities: 1) the customer goes away, perhaps to another supplier; 2) the customer returns at a later time when the Sears stock has been replenished. The former is the *lost sales* case in inventory theory literature; the latter creates a backorder on the supplier. Sears will keep track of customer backorders by logging them and notifying the customer when the item is back in stock. Other retailers will only tell the customer that the item is backordered and that he should reorder after a certain date. In high-technology equipment, any demand that can not be filled is backordered; there are no lost sales and thus we will not treat that case in this book. Like Sears, we will be interested in "supply system performance" measures such as fill rate and number of backorders, but only indirectly. Such measures are used internally in the inventory theory we develop below, but from the point of view of the manager or decision-maker, they are irrelevant. The manager's perspective should be at the system level: What does the optimal *system* cost-effectiveness curve look like? We describe this in the following section.

1.3 The Item Approach

Traditional inventory theory uses the item approach, where the spares for an item are determined by simple formulas that balance the costs of holding inventory, ordering, and stockout. The item approach has been used for years, and it is simpler because decisions on the number of spare units of stock to buy on an item are made without considering other items.

The disadvantage of the item approach is that the availability and total investment in the system of items are uncontrolled outputs of the item decisions. The system availability or investment may be inappropriate. What does the decision-maker do if the item decisions lead to a 35% availability for a fleet of aircraft? Or, if the budget for spares exceeds the money available?

The availability and investment targets should be inputs to the decision-making process. The system approach presents the manager with an availability-cost curve of his 'efficient' system alternatives as illustrated in Figure 1-1. Any points below the curve are "inefficient" in that it is possible to find solutions on the curve with more availability or less cost; points above the curve are unobtainable. The manager chooses the point on the curve that meets the availability requirements within budget limitations. The slope of the availability-cost curve at any point shows the marginal cost of obtaining higher (or lower) availability.

The system approach and the item approach are related in the sense that every point on the system availability-cost curve is computed from an item approach solution for a particular set of parameters: inventory holding cost, order cost, and stockout cost. Thus, to generate the system curve, it is necessary to solve a series of item approach problems. Fortunately there are efficient techniques for generating these curves, and these are described in detail in this book.

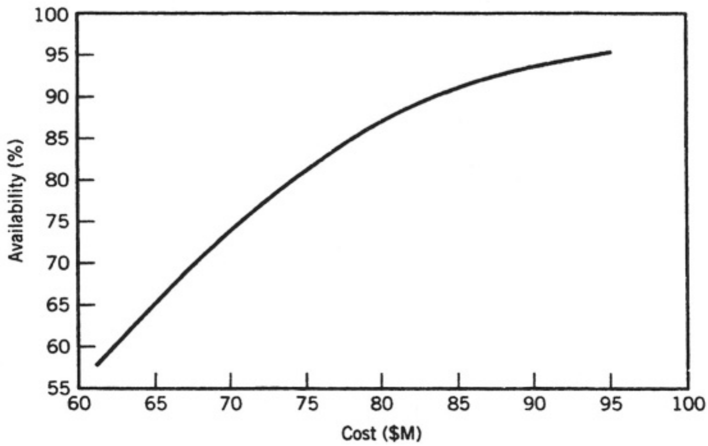


Figure 1-1. The system approach presents the manager with an availability-cost curve of efficient system alternatives.

In 1964 I had the opportunity to visit a military supply depot where all spares were ordered automatically by computer. It was hard to believe that all of the decisions were made automatically and that manufacturer orders were placed from the computer without human intervention. After much probing, the managers finally admitted that the humans had not been replaced completely. "As a matter of fact," they said, "we don't have enough money to do what the computer says, so we buy the projected demand for six months on every item. But, when we get enough money, it will all be automatic." We asked them to call us when that time came, and we're still waiting forty years later. The mismatch between item-level decisions and system resources such as money or system performance requirements does not exist when the system approach is used. Each point on the optimal system cost-effectiveness curve corresponds to a set of stockage policies - a stock level for every item. In the depot experience quoted above, a computer model based on the theory in this book would have obtained better system performance for any spares budget.

1.4 Repairable vs. Consumable Items

Most books on inventory theory begin with consumable or non-repairable items; only later and in a cursory way do they discuss repairable items. They are concerned with two basic questions on each item: (1) when to order, the optimal reorder point (R); (2) how much to order, the order quantity (Q).

Figure 1-2 shows the typical saw-tooth pattern with orders of size Q placed at reorder point R so that the resupply arrives a lead time later, just as the on hand inventory is becoming depleted, as seen in this example where demand is constant and known. (We will generalize this example to probabilistic demand later.)

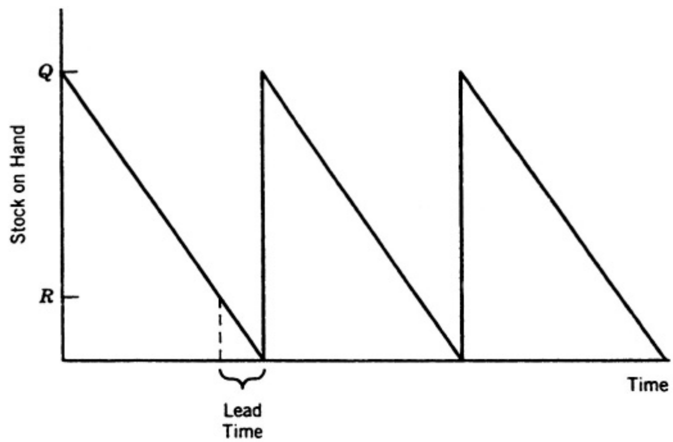


Figure 1-2. **Deterministic demand. *R*, optimal reorder point; *Q*, order quantity.**

This simple example illustrates the fundamental formula of inventory theory, known as the Wilson lot-size formula for the optimal order quantity, which arose in the early part of the twentieth century[1]:

$$Q = \sqrt{\frac{2\Omega m}{\iota c}} \tag{1.1}$$

where Q = the economic order quantity
m = the mean annual demand
$\mathbf{\Omega}$ = the cost to place an order
ι = the annual holding cost rate (e.g. .2 is a common choice for the sum of interest rate, warehousing, and obsolescence)
c = the unit cost of the item

[1] The earliest derivation of this formula appears to be by Ford Harris of the Westinghouse Corporation (1915).

This value of Q minimizes the sum of annual order and holding costs. When we order a lot that has size Q, there are m/Q orders per year and the annual cost is Om/Q. The average amount on hand is $Q/2$ times the unit annual holding cost, ic. The reader is asked to verify Equation 1.1 in Problem 1; it is a special case of a more general formula we derive in Section 5.13.

There are several observations we want to make about Equation 1.1 here. Since our interest is spare parts, and one can't order a fractional number of units of an item, the smallest value of Q is 1. A Q equal to 1 means that we order whenever there is a demand. Our interest is the support of systems, and it turns out that the availability of these is dominated by repairable items. When an aircraft engine develops a malfunction, we don't throw it away - we try to fix it. These repairable items tend to be expensive, and the demand at a base for any particular item tends to be low. As a group the repairable items comprise the largest part of the spares budget; in 1990 the Air Force had over $31 billion invested in repairables. Another reason to pay special attention to repairable spares is that they tend to have longer lead times. If we buy an insufficient quantity, it will take longer to rectify the error.

A small value for m in the numerator of Equation 1.1 and a large value for c in the denominator both tend to make the value of Q approach 1. In effect the repairable item problem has become simpler, because if $Q = 1$ we need to worry about only one "decision variable" on each item - when to reorder. Thus, repairable items are simpler to model than consumable items in this sense; in other respects we will find repairable items are more complicated to model.

As a historical note, the economic order quantity of Equation 1.1 played an important part in our decision to build an optimal spares model. In 1963 Col. Vernon Taylor of the Headquarters USAF staff asked the RAND Corporation to explain why so much attention was paid to unit cost in the EOQ formula used for low-cost items, whereas cost was virtually ignored in the policies for high-cost repairable items. It didn't make much sense to us then, and it still doesn't.

1.5 "Physics" of the Problem

It is important to describe the "physics" of the problem, before we attempt to develop theory. The simplest version of the problem is as follows: When a malfunction is diagnosed on an aircraft, the malfunctioning item is removed from the aircraft and brought into base supply. If a spare is available, it is issued and installed on the aircraft; otherwise a backorder is established for that user. We call this a *first indenture item,* because it is installed directly on the aircraft. Note that a base backorder on a first indenture item implies that there is a “hole” in an aircraft that causes it to be grounded. Later we discuss backorders at the depot and on lower-indenture items at the base. These supply system shortages are important and must be considered in our models, but they do not impact the aircraft directly, e.g. a base backorder for a second indenture item does not necessarily result in a “hole” in an aircraft.

The malfunctioning first indenture item is taken to a base maintenance shop and a determination of repairability is made. If the item can be repaired, it is scheduled into base repair and at some later time, when fixed, it is sent to base supply, where it is used to satisfy an outstanding backorder, if any, or is added to serviceable supply on the shelf. If not base-repairable, it is sent to the depot and a resupply request is levied on the depot. After some resupply delay, the length of which depends on the situation at depot supply, a serviceable unit is received by base supply. Note that usually a different unit of the item is received from the depot than the one sent to the depot. One of the complications of the repairable item theory is that these repair and resupply delays are not fixed; there is a probability distribution for the time to repair an item at the base, depending on the complexity of the repair and the availability of personnel, shop equipment, and spare parts. The order-and-ship time is defined to be the time from placing a request on depot until the time when the item is received at the base *if there was stock on the shelf at the depot.* There is a probability distribution for this time as well. There is also a probability distribution for the waiting time at the depot until an item is available to ship to a base. All of these probability distributions must be taken into account by our theory.

1.6 Multi-Item Optimization

We explained above why the system approach is an important perspective for high-technology equipments. One implication of the system approach is that we will be determining stockage policies on a large number of items. In fact the optimal stockage for different items is not independent of the total number of items. For example, if we want 95% availability on a system with 2000 items, we will need more stock on each item than for a similar system of only 1000 items.

In developing our theory it is important to strike a balance. On the one hand it is important that we not fall into the trap of developing more elegant “exact” solutions that have simplified unduly the “physics” of real-world problems. On the other hand the theory is useful only if it is possible to build efficient computer programs. We have attempted to maintain that balance, and all of the theory in this book has been implemented on personal computers.

1.7 Multi-Echelon Optimization

Another way in which repairable items are more complicated than consumables is that we are typically interested not just in how many spares we need at each operating base, but how many we need at the supporting depot as well. The latter obviously affects the probability distribution that a spare will be on the shelf at the depot to resupply a base. The bases are referred to as the *first echelon,* and the depots as the *second echelon.* The Air Force is considered to be a two-echelon supply system for most purposes. Sometimes there are more echelons. For example, in the support of deployed submarines, some spare stock is kept on each submarine (the first echelon); some is kept on second-echelon supply ships that are periodically accessible by the submarines; these in turn are supported by the third-echelon home port facilities; and finally there are fourth-echelon Navy depots such as Mechanicsburg, Pennsylvania. This multi-echelon picture is more typical of the Army as well. The theory to be developed below is valid for any number of echelons, although the computation takes longer and the computer programs become more complicated as the number of echelons increases.

There is one important assumption in the echelon structure of these models: an "arborescent" or tree structure is assumed wherein each first-echelon base has a specific second-echelon supplier for any given item (the second echelon supplier need not be the same for all items). If there are more echelons, the same type of arborescence is assumed between adjoining echelons, as shown in Figure 1-3. This is an example of "ragged" echelons where the number of echelons may vary; from the viewpoint of the first three bases, there are three echelons whereas from that of the last two, there are only two echelons.

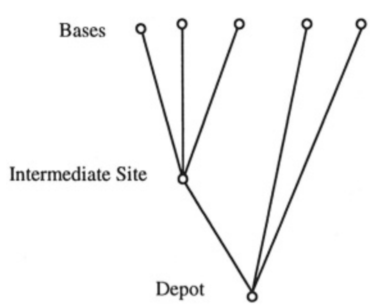


Figure 1-3. **Arborescent tree with "ragged" echelons**

This type of arborescence is not unusual in most inventory systems. But it does imply some operating constraints. For example, suppose a base finds that there is no spare on the shelf at its usual supporting depot; if it is able to go to other depots in search of that spare, it is violating the arborescence assumption. Or if the base can go to other bases to effect a "lateral" shipment, it is violating the assumption.

Some model assumptions are bound to be violated at least occasionally in the real world. The modeler's art is to incorporate in the theory as much of the "physics" of the problem as possible. Thus, we will not prevent lateral shipments from taking place in the real world, but if they become a significant part of the physics they should be included in the theory. (We relax the assumption of no lateral supply in Appendix B).

1.8 Multi-Indenture Optimization

Echelons describe how the supply system is organized. We will also be concerned with the engineering parts hierarchy, referred to as the *indenture structure*. In Air Force terminology a first-indenture item that is removed from the aircraft is called a *line-replaceable unit* (LRU), because this activity takes place on the flight line. When the first-indenture item is taken apart in the maintenance shop, second-indenture items are replaced and these are called *shop-replaceable units* (SRUs). The Navy uses the terms *weapon-replaceable assembly* (WRA) and *shop-replaceable assembly* (SRA) for first- and second-indenture items respectively.

Of course, we can have third-, fourth- and lower-indenture parts as well and a good inventory policy should be concerned with the optimal stockage of these as well. We noted above in Section 1.6 that the optimal stock levels of different items are not independent. This is even more true when we consider trade-offs between items of different indentures. There is a substitution effect because a malfunction of an electronic device (first-indenture item) may be fixed by a circuit card (second-indenture item) or the appropriate computer chip (third-indenture item).

Since an item at a particular indenture is composed of several lower-indenture items, the cost of each lower indenture item is less than that of its "parent". Furthermore, the lower-indenture items, such as computer chips, are more likely to be common items, that is, used in several different "parents". For these reasons, there is an incentive to stock the lower-indenture items rather than their higher-indenture "parents".

On the other hand, when an item fails, it takes time and expertise to diagnose and replace the lower-indenture items that are responsible. It may take specialized test equipment, and it may require sending the item to the depot or next-higher echelon. This extra time translates into system downtime, and for this reason it is preferable to stock the higher-indenture items.

The appropriate mix of different indenture spares, and their optimal allocation between different echelons, is a complicated problem that is best approached with a computer model. That is the subject of this book.

1.9 Field Test Experience

When we use the term "system approach," some logisticians think that all we mean is a different perspective on the problem. In fact, we show in this book that the stockage decisions are significantly different than those that would be made under the traditional item approach.

How different is "significantly different"? We mean the same level of system effectiveness at a spares cost on the order of half as large. It is possible to demonstrate this with computer analyses, and several will be presented in the book. But, there is nothing quite like a field test experiment. If the model has ignored some important physics, it will never be detected by the model itself, but the real-world has ways of educating the careless modeler.

The mathematics of the new approach is the subject of Chapter 2. This approach was tested in several computer simulations and a field test at Hamilton Air Force Base (AFB) where there was one tactical aircraft type. In the six-month test from March-August 1965 the objective was to improve both the performance level of an 82.8% fill rate and lower the investment cost of $1.84 million. The results were a 91.2% fill rate at a cost of $1.45 million. Furthermore, the average number of times per month that an aircraft became not operationally ready for supply dropped from 33 to 19, a 42% reduction.

This gave Air Force decision-makers (and the modelers) greater confidence in the new approach. But there were skeptics who said that only a full field test at a typical, but complex base with several aircraft types would be meaningful. Furthermore, they didn't want a paper test of what would have happened with the stock levels recommended by the model - they wanted a test of what would happen if those levels were actually implemented. Consequently, George AFB with three major aircraft types (F-4C, F-104, F-106) was selected and the stockage system performance under the old policy was monitored for six months by a test team. At the end of this period several freight cars left the base loaded with material that the model said was not needed; several arrived with spares that the new Base Stockage Model (BSM) recommended. Then the test was run for another six months and the performance was compared. The field test was reported on by the Air Force Logistics Command (1967).

The Air Force could have used the model to obtain either better system performance at the same inventory investment, the same system performance at less cost, or some intermediate combination of better performance and less cost as at Hamilton AFB. It chose the second alternative.

In the next chapter we show that backorders are a better performance measure than fill rate, because availability can be computed from backorders. But, in 1965 the Air Force measure of performance was fill rate - the percent of demands that are met when placed. Consequently, that is the measure we use here. Over the group of 3673 repairable items at the base and the stock levels in place during the pre-test period from September 1, 1965 to February 28, 1966 the average fill rate was 75 percent. That was the target that was used in the optimal model for the next six-month period. Table 1-1 compares the results for the pretest and test periods. During the test period from March 1, 1966 to August 31, 1966 the measured fill rate was 76 percent, slightly higher than our target, but the required inventory investment was almost cut in half.

Table 1-1. George AFB Field Test Results

	Pretest	Test	% Change from Pretest
Investment ($M)	13.4	7.3	-46
Fill rate (%)	75	76	1
Special levels	167	28	-83
Aircraft possessed	114	96	-16
Flying hours/month	3621	2264	-37
Sorties/month	2009	1362	-32
Average # of backorders	71	40	-44

How good was the Air Force policy used in the pretest period? In addition to being much more expensive, there were 167 items, 12% of the 1414 items with demand in the first six months, on which the Air Force policy would have stocked an inadequate number of spares in the judgment of experienced personnel. Thus, supply managers had obtained permission for special augmented levels on these items in the pretest period. The stock levels from a model should not be treated like the Ten Commandments. No model embodies all of the day-to-day knowledge of experienced personnel. Thus, there will be occasions when the model output should be adjusted. On the other hand, if it is necessary to revise a substantial amount of the model output on a regular basis, the model needs to be improved.

Thus, we were not surprised that the Air Force felt the model stock levels needed to be increased for 28 items during the second 6-month period. This was, however, a dramatic 83% reduction in the number of special levels that had been used before. Of the 28 items, 9 were in the pre-test group of 167 items with special levels and 19 were high-cost items for which the model had allocated either a zero or low-stock level.

Unfortunately, the antiseptic purity that one would like in a field test was compromised, because of the Vietnam war and the transfer of 16% of the aircraft to Southeast Asia. Even more importantly, the flying hours/month and the sorties/month, the primary factors that influence demand, decreased by averages of 37% and 32%, respectively. The measure most closely related to operational performance is the average number of backorders or aircraft "holes". The actual aircraft availability percentages were not tracked during the test. That is unfortunate, as it would seem to be a better criterion. However, availability percentages are affected by the cannibalization policy. When availability rates fall below targeted levels, maintenance does more cannibalization. The advantage of looking at backorders is that they are unaffected by the cannibalization policy which could change significantly over a test lasting a year.

Although the average backorders declined 44%, more than the sorties or flying hours, we would not claim an improvement in performance. On the other hand, there is no evidence that the 46% reduction in inventory investment has led to a decrease in performance.

Due to the changes in flying hours and sorties between the pretest and test periods, the Air Force simulated what would have happened during the test period under standard Air Force policy and using the Base Stockage Model (BSM). The actual daily demand data and repair times were used in the comparison of the two policies. No special levels were used, resulting in an Air Force (AF) investment of only \$4.3 million. The same budget was allocated by the BSM with the results shown in Table 1-2.

Table 1-2. George AFB Simulation Results

	AF Policy	BSM	% Change
Investment (\$M)	4.3	4.3	0
Fill Rate (%)	53	70	32
Number of Demands	8163	8163	0
Average # of Backorders	61	37	-39

We think the results summarized in Tables 1.1 and 1.2 are very impressive. Had the improvement been 5% or 10%, it might be possible to dismiss it as a chance phenomenon or due to the field test methodology. However, these results have been duplicated elsewhere so they are clearly not a fluke. Furthermore, we remind the reader that these improvements were obtained at a single base. As we consider multi-echelon, multi-indenture systems in the chapters to come, the percentage improvements are usually even larger.

Although these field tests were performed almost 40 years ago, they are still among the best in the sense of actual implementation of the recommended levels and careful monitoring of both the pre-test and test periods. Other tests of the theory are described in the book: 1) Airline results in Section 9.2; 2) Space Station tests in Section 6.13; 3) Demand Prediction Experiments in Appendix C; 4) a C-5 study for the Air Force in Section 10.6; and 5) Coast Guard studies in Section 10.9. The problems of making valid comparisons between the system approach developed in this book and any other procedure for stock leveling are discussed in Section 10.2.

1.10 The Item Approach Revisited

How did we achieve such impressive results in the field test? In the next chapter we develop the theory for the single-site model. Chapter 4 describes the Bayesian analysis procedures that were used to obtain better estimates of demand. In this section we describe the policy then used by the Air Force. The policy is still used for some items by the Defense Logistics Agency and others, and can be characterized as *the item approach,* which has been used traditionally to determine the required number of spare parts. We begin with a single item at a base where we assume that the item can always be repaired. Let the average demand over the lead time be denoted as μ; where in this case the lead time equals the repair time. Then the traditional policy has been to buy enough spares to cover the lead time demand plus some safety level to protect against demand variability:

$$s = \mu + k\sigma \tag{1.2}$$

where s = units of spare stock
μ = average demand over the lead time
σ = standard deviation of lead time demand
k = positive constant for the amount of protection

The standard deviation of lead time demand is the square root of μ, when the lead time itself is constant and demand has a Poisson distribution. The Poisson will be described in detail in the next chapter, but it is the common choice for modeling random demand, as contrasted with wear-out phenomena. A typical value of k used by the Air Force is the square root of 3 = 1.73. (The values on the right-hand side of Equation 1.2 may be non-integers, so the result is rounded to obtain an integer value of s).

It turns out that the protection level, k, should not be a constant across all items. We will prove this formally in Chapter 2, but it is important that the reader understand why a variable protection level makes sense. Suppose there are two items that have exactly the same demand characteristics and each is equally important to the operation of the aircraft. One costs $100 and the other $10,000. If we are attempting to allocate a fixed budget to obtain the highest availability possible for a fleet of aircraft, we should buy a little more of the inexpensive item. An additional unit of either item produces the same increase in availability, but the expensive item costs one hundred times as much as the inexpensive item. Twenty-five years ago when we were first using this argument, there was a lot of resistance. The logistician would say that if he needs the item, it doesn't matter how much it costs - it is still cheaper than an aircraft. He is right. When a "hole" in an aircraft occurs, we do anything to fill it; expedite maintenance, priority shipment, lateral supply, cannibalization. But that is after the failure occurs. Our problem is like placing bets on an upcoming horse race; we want to allocate our spares budget across a group of items *before* any failures occur, and at that point in time the unit cost of the item should be considered. Thus, k should be smaller for high-cost items, other things being equal. We will provide an example in the next section.

In Chapter 3 we will discuss multi-echelon stockage and in Sections 4.4 and 4.5, multi-indenture stockage. It turns out that the optimal value for k should be different by echelon. Not surprisingly, the first-echelon or base level, where aircraft are flown, is the most important. By contrast, the stock level at the depot affects the time to resupply the bases, but the impact on aircraft availability at the bases is indirect. It will turn out that the k value should be larger for the base than for the depot. Similarly, the first-indenture items at the base have a more direct impact on aircraft availability than second-, third-, or lower-indenture component parts. Again the optimal k should decrease as we move down the indenture structure.

All of this may seem very complex, but in fact we need never worry about those k values. They are implied by the mathematics, but never computed directly. We need, however, to understand why k should not be a constant across all items, echelons, and indentures for a point on the optimal cost-effectiveness curve. This in turn helps to explain why simple inventory policies with a constant protection level, k, on all items, such as the policy at George AFB in 1965, produce dramatically inferior results.

1.11 The System Approach Revisited

In this section we contrast the results from the item approach and the system approach when there are a small number of items. To facilitate comparison we assume there are 22 first-indenture items, each of which is critical to the operation of the aircraft. Item demand is Poisson with costs and demand rates as given in Table 1-3. We assume that each item can be repaired in an average of 10 days and consider a single base.

Table 1-3. Example Data, Section 1.11

Item #	Cost	Average Demands/Day	Average Pipeline
1	$1000	0.1	1
2-11	$ 100	0.1	1
12	$1000	1.0	10
13-22	$ 100	1.0	10

We will use the term "pipeline" throughout the book to denote the number of units of an item in repair at a site or being resupplied to the site from a higher echelon. The number of units in the pipeline varies probabilistically, but the pipeline can be measured at any point in time by counting the number of units in repair and resupply. The average pipeline for a site or the average lead time demand is the average number of units in repair/resupply.

Suppose that we have a fleet of 100 aircraft and our objective is to maximize the availability of the fleet. Assume that a spares budget of $22,000 is allocated across the items using: a) the item approach of Equation 1.2; b) the maximization of availability when there is *no* cannibalization; and c) the maximization of availability when there is cannibalization. We want to look at both the system performance and the underlying stock levels for each method. The item demands and costs have been chosen purposely to facilitate our comparisons.

The budget of $22,000 is only enough to buy the average pipeline on every item using the item approach of Equation 1.2. This is called the Constant k (Const. k) policy in Table 1-4, where $k = 0$. Using the methods to be derived in Chapter 2, we obtain the optimal stock levels for no cannibalization in the next column of Table 1-4 (No Cann.). Finally, using the theory for cannibalization from Chapter 8, the last column with optimal cannibalization stock levels (Cann.) is obtained.

Table 1-4. Optimal Policies for the Example

Item	Cost	Demands/ Day	Average Pipeline	Const. k	Optimized No Cann.	Optimized Cann.
1	$1000	0.1	1	1	0	0
2-11	$ 100	0.1	1	1	2	0
12	$1000	1.0	10	10	6	9
13-22	$ 100	1.0	10	10	14	13
Evaluated Availability						
No. Cann				83.61%	92.21%	85.13%
Cann.				94.64%	95.34%	96.04%

One of the advantages of the system approach is that we can *evaluate* the availability of any set of stock levels. Thus, the model of Chapter 2 is used not only to find the optimal allocation of $22,000 and the corresponding availability of 92.21% (assuming a fleet of 100 aircraft). It is used to evaluate the availability achieved with the Constant *k* and Cannibalization policies, also. Note that the Constant *k* policy yields an availability of 83.61% if no cannibalization is practiced, and the policy optimized under the assumption of cannibalization yields 85.13% if no cannibalization is practiced. The policy that is optimized under the assumption of no cannibalization must have the highest availability when evaluated under the same assumption. That is why we have underlined the 92.21% availability as the optimum in the first row of evaluations with no cannibalization in Table 1-4.

Similarly when the cannibalization model of Chapter 8 is used to evaluate the three policies in the last row of Table 1-4 (Availability with cannibalization), the policy that was optimized under the same assumption must be best. We have underlined 96.04% for this reason.

What can we conclude at the system level? 1) The Constant *k* approach is inferior to both optimal policies, regardless of whether or not maintenance cannibalizes; 2) If maintenance does not cannibalize, we should plan for that and optimize assuming no cannibalization; 92.21% is significantly higher than 83.61% and 85.13% in the same row; 3) If maintenance does cannibalize, our availability will be slightly higher if we optimize accordingly. But, the best availability of 96.04% is not dramatically larger than either 94.64% or 95.34% in the last row.

Of course, we cannot make sweeping conclusions from a single example, but it turns out that this phenomenon is often observed: stockage policies that are optimized for no-cannibalization tend to perform well even if maintenance does cannibalize, but not vice versa. To understand why, we must examine the stock levels.

Note that the stock levels in all three policies are most strongly affected by the demand rates. For a specific cost, the demand rate has the greatest impact on the Constant *k* policy - a factor of ten in the demand rate translates to a factor of ten in the stock level. For a specific demand rate, a factor of ten in the cost translates to a factor of two or less in the stock levels for the no-cannibalization and cannibalization policies. Comparing these two optimal policies, we see that demand rate is more important in the latter case. This is because we assume that we can consolidate "holes" with cannibalization, and therefore depth of stockage on high demand items is more important than breadth.

Before leaving this topic we should recall that Poisson demand was assumed in Table 1-4. In later chapters, we consider the problems of demand rates that do not stay constant, but "drift" up or down over time. This leads to more variability in the demand process, and greater differences in the availabilities and stockage policies. Table 1-5 shows the results when demand has a negative binomial distribution, and variance-to-mean ratios that increase with the mean, as described in Chapter 4.

All of the availabilities in Table 1-5 are much lower, and the Constant k policy looks even worse than before. Suppose that maintenance does decide to practice cannibalization and the stockage policy was optimized under the assumption of *no* cannibalization. There is a modest degradation from the 91.63% availability that could have been achieved to 90.65%. Thus, the *no* cannibalization policy is quite "robust." However, if maintenance does not cannibalize and the stockage policy was optimized under the assumption of cannibalization, there is a larger degradation, from the 84.62% availability that could have been achieved to 79.90%. The cannibalization policy is less robust. Of course, the differences between availabilities will increase dramatically as we move from 22 items to the hundreds or thousands of items that compose a typical complex system.

Table 1-5. Optimal Policies for Negative Binomial Demand

Item #	Cost	Demand/ Day	Average Pipeline	Var. /Mn	Const. k	Optimum No Cann.	Optimum Cann.
1	$1000	0.1	1	1.85	1	0	0
2-11	$ 100	0.1	1	1.85	1	3	0
12	$1000	1.0	10	3.67	10	3	5
13-22	$ 100	1.0	10	3.67	10	16	17
Evaluated Availability							
No Cann.					67.93%	84.62%	79.90%
Cann.					86.74%	90.65%	91.63%

The cannibalization and no cannibalization policies in Table 1-5 are more sensitive to item cost than the corresponding policies in Table 1-4; both are about equally sensitive to differences in demand rates.

1.12 Summary

The system approach is superior to the item approach for managing support of equipments. Not only does the system approach provide management with some assurance about the availability levels that should be attained, but any specified availability is achieved at dramatically lower investment. This has been demonstrated repeatedly both in computer simulations and actual field tests over a period of 40 years.

We have seen that the explanation for these improvements is that the protection level, k, should be a variable. It depends on unit cost, echelon, and indenture. It also depends on whether cannibalization is practiced or not. However, we need never calculate the k values, because the optimal stockage policies are determined directly from the algorithms derived in this book.

1.13 Problems

1. In the case of known, constant demand discussed in Section 1.4, write an expression for the total annual costs due to ordering and holding inventory. Find the minimum cost by differentiating with respect to Q (the economic order quantity) and setting the result to zero, thus deriving Equation 1.1 for the optimal Q. Plot the two cost terms (annual costs of ordering and annual costs of holding inventory) as functions of Q; then plot the sum of the two costs, the total cost. What is the relationship of the two cost terms at the optimal value of Q? Suppose that $\iota = .2$, c (the cost of the item) = \$1500, and m (the mean annual demand) = 1. Since any optimal Q up to 1.49 will round to 1, use that Q to determine the maximum value of order cost, Ω, under which the optimal policy is one-for-one replenishment. Use several larger values of m as well. (See Problem 12 of Chapter 2 for the impact of probabilistic demand).

2. The Constant k policy in Table 1-4 corresponds to a $k = 0$. Show that the k values for the no-cannibalization policy are -1, 1, -1.26, and 1.26 for the four groups of items. Thus the protection levels are higher for the lower-cost items. Find the implied k values for the cannibalization policy in Table 1-4.

Chapter 2

SINGLE-SITE INVENTORY MODEL FOR REPAIRABLE ITEMS

It is better to know nothing than to know what ain't so.

-Josh Billings

2.1 Chapter Overview

In this chapter we develop the basic model for a single operating base. Our objective is to develop a curve showing system availability for a fleet of aircraft as a function of optimal spares investment over a group of items. Our assumption here is that each item is equally important, and that a backorder for any item is equally serious in that an aircraft is grounded. Remember that a base backorder implies there is a "hole" in an aircraft when the items needed are all first-indenture items. We assume that cannibalization is not performed to consolidate backorders onto the smallest number of aircraft. (In Chapter 4 we relax the assumptions that all items are first-indenture and in Chapter 8 cannibalization is considered.)

We begin by presenting formulas for the mean and variance of any probability distribution. Then the Poisson probability distribution is introduced. This is followed by Palm's theorem, which is crucial to estimating the probability distribution for the number of units of an item in repair at a random point in time. We define expected fill rate and expected backorders, two measures of item performance, and show how these can be calculated from the probability distribution of the number of units in repair. This material was originally presented in Feeney and Sherbrooke (1966). System performance measures such as availability are defined also. Then we develop the single-site model and show how to compute an optimal curve relating investment cost to expected system backorders. It is shown that the maximization of availability is obtained by the minimization of expected system backorders using a derivation by Smith, Fisher, Heller (1972). Thus, the optimal curve for investment cost versus expected backorders can be transformed immediately into an optimal curve for investment cost versus expected system availability. Each point on the optimal system availability-cost curve corresponds to a set of stockage policies - a stock level for every item.

2.2 Mean and Variance

Let X be a random variable and $\Pr\{X = x\}$ designate the probability that the random variable X takes on a specific value x from some unspecified probability distribution. A requirement for a probability distribution is that $\Pr\{X = x\}$ is non-negative for any x and the sum of probabilities over all x equals one. In this book we shall be concerned primarily with probability distributions where X is a discrete random variable taking on values 0, 1, 2... We use the notation $E[X]$ to represent the *expected value* or *mean*[1] of the random variable X, and it is defined as:

$$E[X] = \sum_{x=1}^{\infty} x \Pr\{X = x\} \tag{2.1}$$

We will need a measure of the spread of the X values around their mean as well. The expected value of this dispersion around the mean is called the *variance*:

$$Var[X] = E[X - E[X]]^2 = E[X^2] - (E[X])^2 \tag{2.2}$$

[1] We use the term *expected* to indicate the probabilistic weighting of all possible outcomes; the *expected value* of a quantity is also called the *mean*. The term *average* is used for an observed or measured quantity, but it is also a synonym for mean.

where the first term on the right-hand-side, $E[X^2]$, is the expected value of X^2, also known as the second moment of X, defined similarly to Equation 2.1 as

$$E[X^2] = \sum_{x=1}^{\infty} x^2 \Pr\{X = x\} \tag{2.3}$$

and the second term is just the square of the mean from Equation 2.1. From the definition of the variance in Equation 2.2 as a sum of squares, it is easy to see that the variance is always non-negative for any probability distribution.

2.3 Poisson Distribution and Notation

The Poisson distribution, $p(x)$, is given by:

$$p(x) = (mT)^x e^{-mT} / x! \qquad x = 0, 1, 2 \ldots \tag{2.4}$$

where the mean, $E[X]$, from Equation 2.1 is found to equal mT. We will find it convenient in this book to define m as the average annual demand and T as the average time period measured in years. Of course, the Poisson distribution is unaffected by the time unit used for m and T, because the mean depends only on their product. The variance, Var[X], from Equation 2.2 is mT, also. It is shown in Section 4.1 that when the time between demands is given by an exponential distribution (also called a *Poisson process*), the number of demands in a time period of any fixed length is given by the Poisson distribution. The exponential distribution is the "memoryless" distribution in which the time of the last demand has no influence on the time of the next demand. Since random failures are the primary type for which our models are designed, we shall use the Poisson extensively. Later we will consider Poisson processes whose mean changes over time[1], as well as items whose failure is related to wear-out phenomena.

[1] This is called a *Poisson process with non-stationary increments* or a non-homogeneous Poisson process in the literature. We will use the former term to indicate cases where the mean changes; when we refer to a Poisson process without qualification, this indicates a process with *stationary increments.* A process refers to how something evolves over time. We also use the term *state probabilities* to refer to the number of demands over a specified period of time (states of 0, 1, 2, . . . etc.). There is some confusion because a Poisson process leads to Poisson state probabilities; a Poisson process with non-stationary increments leads to state probabilities that can be *approximated* by a negative binomial distribution, as shown in Section 4.8.

Because the Poisson distribution is fundamental to much of our analysis throughout the book, we use the mnemonic $p(x)$ for the Poisson probability density function. When we need to differentiate between Poisson distributions with different means, we use the notation $p(x|mT)$ to indicate that the Poisson distribution (or other probability distributions to be defined) is "conditional" on the parameters to the right of the vertical bar. A complete list of variables is provided on pages xv-xviii.

2.4 Palm's Theorem

The cornerstone of repairable item inventory theory is a queueing theorem of Palm's (1938). Its importance is that it enables us to estimate the steady-state probability distribution of the number of units in repair from the probability distribution of the demand process and the mean of the repair time distribution.

> PALM'S THEOREM. If demand for an item is a Poisson process with annual mean m and the repair time for each failed unit is independently and identically distributed according to any distribution with mean T years, then the steady-state probability distribution for the number of units in repair has a Poisson distribution with mean mT.

This is sometimes called the *infinite channel queueing assumption,* because there is really no queueing or interaction in the repair times of the several items. Nevertheless, the theorem is remarkable in that it is unnecessary to measure the shape of the repair distribution. For any specified mean time T, regardless of the distribution, the steady-state probability distribution for the number of items in repair is Poisson with mean mT. The theorem is proved in Appendix A.

From the logistician's viewpoint this is an extremely important result, because there is no need to collect data on the shapes of the repair distributions.

2.5 Justification of Independent Repair Times and Constant Demand

How can we justify the modeling assumption that there is no interaction in the repair times of the several items? Obviously the modeling is simplified by this assumption because it lets us use Palm's theorem, but is this reasonable?

Let's consider the "physics" of the repair process. When an item is brought into the repair shop, a common procedure is to take it to a test stand for diagnosis. The test stand and the repairman are capable of handling a number of different items, so that if we really want to measure the repair queueing that may take place, we must consider the entire group of items that compete for the test stand and the repairman. The problem is complicated by the fact that different types of failures on a given item may require a different test stand/shop and a repairman with different skills. Was it an electrical system failure? Or was it related to a mechanical problem such as a broken connector pin? Perhaps the failure is in the hydraulics system rather than in the electrical system.

The problem is further complicated by the fact that items are not necessarily fixed on a first-come, first-served basis. An item that is grounding an aircraft is going to be put at the head of the repair line. In effect, we need to model the repair shop management process as well. Thus, if we want to model the shop in detail, we have a very complex problem.

Even if we can't model the repair queueing exactly, can we determine whether our assumptions of independence of repair times lead to optimistic or pessimistic results? In the real world some queueing does take place, and our model would understate the real repair delay. On the other hand in those cases of greatest interest where there is a "hole" in an aircraft for the item, maintenance management is going to expedite repair and our model will overstate the repair delay.

While these phenomena tend to offset each other, it is probable that the net result of the independent repair time assumption is to understate backorders. Simulation is of limited utility in estimating the error, because it is difficult to replicate management behavior in a computer. Furthermore, different managers will behave differently. In fact our model does include delay, because the actual man-hours spent repairing an item are usually a small fraction of the average repair time. Most of the time the item is waiting for something: parts, test equipment, maintenance personnel. What we have not modeled are the detailed queueing interactions. While it might be nice to have a more sophisticated model, we must keep in mind that the data are estimated values with substantial error - not physical constants like the speed of light which can be measured with great precision. We believe that the assumption of independent repair times is a reasonable approximation, and this is reinforced by over forty years of usage.

Gross (1982) does model the queueing process for finite servers when repair times have an exponential distribution. He derives expressions for expected fill rate and expected backorders, and makes comparisons with the infinite channel assumption. His theory will be more accurate than ours in a case where there are a fixed number of test stands, and they are dedicated to a particular item. In effect the finite server model assumes there is no need to model maintenance management, because the only delays are for test equipment. This implies that once an item is on the test stand, it stays until it is repaired, regardless of the type of failure or broken parts. These assumptions and the restriction to exponential repair times seem less realistic in our application than that of independent repair times. For these reasons we will not embed the Gross theory in our modeling below.

Another important assumption of Palm's theorem is that the average annual demand, m, stays constant. But, when aircraft are grounded, the real demand rate decreases; for example, if 25% of the aircraft are grounded, the demand rate will be 25% less. Although this is a mathematical difficulty, it does not affect the logistics application significantly. This is because we will be interested in stockage policies that produce high availabilities, such as 90%, for the fleet of aircraft. Thus, the error in assuming a constant demand rate will be on the order of 10%, well within the accuracy limitations of logistics data.

The assumption of a constant demand rate will tend to overstate backorders, tending to counteract to some extent the independent repair time assumption which probably understates backorders.

2.6 Stock Level

Our inventory theory objective is to compute optimal stock levels for each item. What is a stock level, anyway? It can be thought of as the total number of spare units of the item that we want to procure in initial provisioning. Consider one item at a single base where we assume each failed unit of the item can be repaired in a time drawn from a probability distribution with mean T. Assume for the moment that the item can always be fixed (no condemnations).

How does the stock level, s, relate to the quantity on the shelf at the base? There will be times when all s spares are on the shelf in good condition. At other times there will be units undergoing repair, which results in less than s on the shelf. There may be times when there is nothing on the shelf, and we have outstanding backorders for customers who could not be satisfied. At these times we will have even more than s in repair. (Sometimes the stock clerk with limited experience is misled into thinking that a stock level of s indicates that there should always be s units on the shelf).

Since all s spare assets must be somewhere, we can write a stock balance equation that is the basis for all of our analysis to come.[1]

$$s = \mathrm{OH} + \mathrm{DI} - \mathrm{BO} \tag{2.5}$$

[1] The quantity s in Equation 2.5 is called the *inventory position* in many texts. The inventory position is not a constant unless the order quantity, Q, equals one.

The stock level, s, is a constant when the order quantity is one and the reorder point is s - 1. The number of units of stock on hand, (OH), the number of units of stock *due in* from repair and resupply, as generalized below, (DI), and the number of backorders, (BO), are non-negative random variables. Any change in one of these random variables is accompanied by a simultaneous change in another. For example, when a demand occurs, the number due in from repair increases by one. If the stock on hand is positive, it is decreased by one; otherwise, the backorders increase by one. In either case the equality is maintained. When a repair is completed, reducing DI by one, the backorders are reduced by one or, if there are no backorders, the on hand balance is increased by one. Again the equality is preserved.

The simplicity of Equation 2.5 is due to the fact that the economic order quantity, Q, for the batch size to repair equals one. This is because these repairable items tend to be high-cost, and low demand at a base as discussed in Chapter 1. Because of this one-for-one repair, the reorder point (or the asset position at which we send an item to repair) is s - 1. In this chapter we assume that all units of each item are repairable at base, and thus the due-ins equal the number of units in base repair; later when we generalize to multi-echelon, the due ins are the sum of units in base repair and those in resupply from the depot, but Equation 2.5 is still valid.

The literature on inventory theory sometimes uses the notation (s, s) inventory policy to indicate that when the inventory position (on hand plus on order minus backorders) drops to s, an order should be placed for s - s units. That is why the repairable item inventory policy is called (s-1, s).

Equation 2.5 is critical to the theory to be developed. We will derive the probability distribution for the number of units in repair. That knowledge in combination with a specified stock level, s, determines both the probability distribution for stock on hand and the probability distribution for backorders. We are able to determine both distributions, because when the number in repair is less than s there is stock on hand; when it is more than s, there are backorders.

2.7 Item Performance Measures

As mentioned in Chapter 1, there are two principal measures of item performance: *fill rate,* the percentage of demands that can be met at the time they are placed; and *backorders,* the number of unfilled demands that exist at a point in time. Whenever we are unable to fill a demand, a backorder is established. The backorder lasts until there is a resupply or a failed item is repaired (a due in, DI, is satisfied). These two measures are related, but differ in very important ways. Fill rate is concerned only with what happens at the time a demand is placed. Backorders measure the number of demands that have not been satisfied at any point in time.

Our interest is in the expected values of these measures. Thus we want to estimate these quantities from the stock level, *s,* and the steady-state probabilities for the number due in (in repair or resupply). Regarding fill rate, there will be a fill if the number due in is $s - 1$ or less, because that implies there is stock on hand. Whenever the number due in is *s* or more, there is no stock on hand. Thus, if we designate the expected fill rate as EFR(*s*):

$$\begin{aligned} \mathrm{EFR}(s) &= \Pr\{\mathrm{DI}=0\} + \Pr\{\mathrm{DI}=1\} + \ldots \Pr\{\mathrm{DI}=s\text{-}1\} \\ &= \Pr\{\mathrm{DI} \le s\text{-}1\} \end{aligned} \tag{2.6}$$

where the Pr{ } terms are the steady-state probabilities for the number of units of stock due in. (For the time being, think of these as Poisson probabilities with mean *mT* as given by Palm's theorem. Later we use other probability distributions as well.) Note that if $s = 0$, the expected fill rate is zero. As *s* increases the fill rate approaches one. Typically the expected fill rate is multiplied by 100 and given as a percentage.

A similar exercise will be used to estimate the expected backorders.[1] Suppose that there are $s + k$ units of stock due in at a random point in time; then from Equation 2.5 there are *k* backorders. The expected number of backorders, EBO(*s*), is thus:

$$\begin{aligned} \mathrm{EBO}(s) &= \Pr\{\mathrm{DI}=s+1\} + 2\Pr\{\mathrm{DI}=s+2\} + 3\Pr\{\mathrm{DI}=s+3\} + \ldots \\ &= \sum_{x=s+1}^{\infty} (x-s)\Pr\{\mathrm{DI}=x\} \end{aligned} \tag{2.7}$$

The expected number of backorders is a non-negative quantity. Note that when $s = 0$, Equation 2.7 becomes identical to Equation 2.1 for the mean of a distribution. Thus, EBO(0) = $E[X]$.

[1] The number of backorders is defined as: $B(X|s) = (X - s)$ if $X > s$
$= 0$ otherwise

where *X* is the random variable for the number of units due in and *s* is the stock level. We are concerned only with the expected number of backorders, where the expectation is taken over the variable *X,* $E[B(X|s)]$, as given by Equation 2.7. Since this is no longer a function of *X,* we use EBO(*s*) to simplify the notation in the text. We used a similar simplification for expected fill rate, EFR(*s*) and later for the variance in backorders, VBO(*s*).

It is important to appreciate the differences between fill rate and backorders. If the stock level on an item is increased, the fill rate will increase and the backorders decrease, but it is possible to have a system with a low fill rate and low backorders. This could happen if the stock levels are low, but the repair facility is very responsive.

The converse situation of a high fill rate and high backorders doesn't arise. If backorders are high on average, then stock on hand will tend to be low, and a low on hand balance produces a low fill rate.

Fill rate measures only what happens when demands occur, whereas backorders measure the duration of shortages. A comparative example over a period of a year is shown in Figure 2-1. Fill rate is the number of downward transitions from a positive on hand to a positive or zero on hand (fills) divided by the total number of downward transitions (demands). In Figure 2-1 there was a unit of stock on the shelf when the first demand occurred at .15 years, resulting in a fill. There was nothing on the shelf at .25 years when the second demand occurred, and one backorder at .5 years when the third demand occurred. Since the last two demands could not be filled, the average fill rate over the year was 33%. The average backorders of 1 are computed from the shaded (negative) area, divided by the length of time, in this case one year. (The term "average" is used rather than "expected" in Figure 2-1, because these are computed over a year rather than over the probability distribution of demand. As the length of time increases, the average becomes a better estimator for the expected value.)

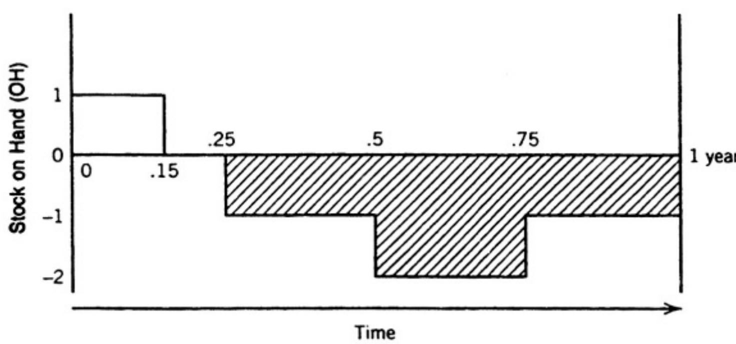


Figure 2-1. **Example of fill rate and backorders over one year. The average fill rate = 33%; the average backorders = 1.**

In fact the average backorders can be calculated in either of two equivalent ways: (1) consider a long period of time, add up the number of days that each backorder has been unfilled, and divide it by the total length of the time period under consideration (as in Figure 2-1); (2) pick a number of different times at random and average the number of backorders observed at these times. Note that the backorder measure weights one backorder of t-days duration as equivalent to t backorders each lasting a day.

Inventory managers tend to like fill rate, because it is easy to measure. They need only consider what happens at the time each demand is placed - was it filled or not. Furthermore, the numerical value has meaning to them. They may know that a 95% fill rate is acceptable in their application, but lower values result in a lot of customer complaints.

Backorders are harder to compute, because it is necessary to keep track of the number of customers who still have a shortage (imagine Sears trying to estimate this quantity). The numerical value is less meaningful to the manager, e.g. how satisfactory is an average of .1 backorders per item?

Even though backorders tend to be less intuitive for the inventory manager, they turn out to be more useful to us because system availability is maximized when the sum of backorders across items is minimized. This is shown below in Section 2.14. From the perspective of the system manager who is trying to fly airplanes, availability is the most meaningful measure. The fact that the model uses backorders internally in the calculation is irrelevant to the system manager. Though we minimize backorders in the process of maximizing availability, we can and will calculate fill rate also, so that inventory managers can monitor inventory performance with a more familiar measure.

Are there other item measures of interest to us? The terms *service rate, ready rate,* and *probability of sufficiency* are sometimes encountered. Service rate is usually the fill rate, but it some applications it measures something else; we will not employ the term in this book, because of its ambiguity. We defined the term *ready rate* in Feeney and Sherbrooke (1966) to be the probability that an item observed at a random point in time has no backorders. Equation 2.6 is the computational formula, except that the summation is from 0 to s. When $s = 0$ the ready rate is positive, even though the fill rate is zero. Since the ready rate is seldom used as a logistics measure, we will not employ it below.

The *probability of sufficiency* is mathematically identical to ready rate, and we do use it in Section 6.5, but in a context where there is no repair and resupply is periodic. In that situation, the probabilities are calculated for the number of demands during the resupply cycle, not for the steady-state probabilities for the number of units in repair. (Note that expected fill rate or expected backorders could be used as measures in a problem with no repair, also. Equations 2.6 and 2.7 respectively would be correct if demand probabilities are used.) If demand is Poisson and there is no repair, the mean is computed over the resupply cycle time. If demand is Poisson and Palm's theorem applies to the repair process, the mean is computed over the average repair time.

2.8 System Performance Measures

Our primary measure of system level performance is *availability,* the expected percentage of a fleet of aircraft that is not down for spares at a random point in time. If there is only one aircraft or end item, the availability is the percent of time that the aircraft is operational. As in the item measures of performance, we will be interested in the expected value of availability in the steady-state where it does not vary. In Chapter 6 we will consider an application to the space station, where resupply is periodic and availability declines between resupply missions. Even though the availability is different at each time t in the cycle, the availability we calculate for a specific value of t is an expected availability in the sense that it is the percent of cycles where the station is not down for spares at time t.

Another measure of system performance that is sometimes used is the *probability of y or fewer aircraft down,* where y is a specified parameter. The probability of y or fewer aircraft down is more concerned with the ability to do specific things with the aircraft than with a general capability. For example, it has been used to a great extent in wartime scenarios where the operational planners have a fleet of N aircraft and they want to be able to launch a wave with at least $N - y$ of the fleet. This performance measure is particularly convenient analytically in applications where cannibalization is practiced, as discussed in Chapter 8. This has sometimes been called *the confidence level* approach or the *direct support objective* (DSO). We will not use these terms.

2.9 Single-Site Model

We are now ready to develop the theory for a single site to obtain the optimal curve for system availability versus investment cost. This is done by developing an optimal curve for total backorders versus system cost, and then showing that minimization of the sum of the backorders on all items is equivalent to maximizing availability.

To simplify the exposition we start with an example comprising two items, both with Poisson demand as defined in Equation 2.4. The items are shown in Table 2-1, together with the expected backorders for various stock levels, as computed from Equation 2.7.

The term *pipeline* was introduced in Chapter 1. It is a measurable quantity, the number of units of the item in repair. Note that m is the average annual demand and T is the average repair time in years, so that the average pipeline, μ, is dimensionless and $\mu = mT$. Because of Palm's theorem, we know that the average pipeline is the mean of the Poisson distribution used to calculate the expected backorders.

Table 2-1. Numerical Example for Single-Site Model

Item		1	2
Average annual demand (m)		10	50
Average repair time yrs. (T)		.1	.08
Average pipeline (μ= mT)		1	4
Item cost (000)		\$5	\$1
	s	EBO(s)	EBO(s)
	0	1.000	4.000
	1	.368	3.018
	2	.104	2.110
	3	.023	1.348
	4	.004	.782
	5	.001	.410
	6	.000	.195
	7	.000	.085
	8	.000	.034
	9	.000	.012
	10	.000	.004

Let's begin by computing a single point on the optimal backorder-versus-cost curve. Suppose that this should be .2 or fewer total backorders. By trial and error we can find a couple of admissible combinations of stock levels on item 1 and item 2 as is shown in Table 2-2.

Table 2-2. Trial-and-Error Solution

	Item 1			Item 2		System	
s	EBO(s)	Cost(\$000)	s	EBO(s)	Cost(\$000)	EBO(s)	Cost(\$000)
2	.104	10	7	.085	7	.189	17
4	.004	20	6	.195	6	.199	26

The upper line in Table 2-2 is a better solution, because it has fewer backorders and costs less. It is easy in this case to verify by trial and error that there is no cheaper combination of the two items that has system backorders of .2 or less.

2.10 Marginal Analysis

The trial-and-error procedure is not an efficient way to develop an optimal backorder-versus-cost curve. Instead we will use a technique called *marginal analysis,* and prove that it produces an optimal curve. The technique is called marginal analysis because at each step in the algorithm we need look only at one number for each item to determine the next item that should be bought. The marginal or incremental value provides all the information necessary on each item. While it is likely that the technique has been used for many years, the earliest published reference appears to be O. Gross (1956).

Sometimes a name impedes communication. The late Murray Geisler, an active logistician for many years, tells the story of briefing an Air Force general in the Strategic Air Command (SAC) about the marginal analysis technique. The general objected to the technique saying that SAC would never accept anything that was "marginal". Recent references sometimes refer to the technique as the "greedy heuristic", hardly an improvement.

We will illustrate the use of marginal analysis, and then justify its optimality. Table 2-3 repeats the expected backorder columns from Table 2-1 plus an additional column for each item - the marginal decrease in expected backorders divided by the item cost (in thousands of dollars).

$$[\mathrm{EBO}(s-1) - \mathrm{EBO}(s)]/c$$

This is the increase in system effectiveness per dollar ("bang per buck"), obtained when an additional unit of that item is selected for stockage. We will call it the *delta value* to emphasize that it measures the change.

The marginal analysis technique for generating the optimal system backorder-versus-cost curve starts at the top number in each of the delta columns and selects the item with the larger value. It then moves to the next lower value in the delta column of the selected item and compares that with the original delta value for the other item. At each step we buy another unit of that item with the larger value, and move to the next lower value from that column for the next comparison. The total system backorders at each point is the sum of the two backorder values next to the last delta values selected.

Table 2-3. Marginal Analysis

	Item 1		Item 2	
s	EBO(s)	[EBO(s - 1) – EBO(s)]/5	EBO(s)	[EBO(s - 1) – EBO(s)]/1
0	1.000		4.000	
1	.368	.126*	3.018	.982*
2	.104	.053	2.110	.908
3	.023	.016	1.348	.762
4	.004	.004	.782	.567
5	.001		.410	.371
6	.000		.195	.215
7	.000		.085	.111
8	.000		.034	.051
9	.000		.012	.021
10	.000		.004	.008

***These columns are referred to as the *delta columns* in the text.**

Thus, in Table 2-3 the total expected backorders EBO(s) before any stock is purchased are 1.000 + 4.000 = 5.000. The deltas for the first spare of item

1 and item 2 are .126 and .982 respectively, so we select item 2 and our total backorders drop to 1.000 + 3.018 = 4.018. We move down to the second spare of item 2 and now compare deltas of .126 and .908. Again item 2 wins and after adding the second spare of item 2 the total backorders are now 1.000 + 2.110 = 3.110. The first six comparisons all result in the selection of item 2, after which the delta of .126 for item 1 exceeds .111 for the seventh spare of item 2, so the first unit of item 1 is added next. The result is the system backorder versus cost curve as displayed in Figure 2-2. Note that at a cost of $17,000 the total backorders of .189 agree with the trial-and-error solution that we found in Table 2-2.

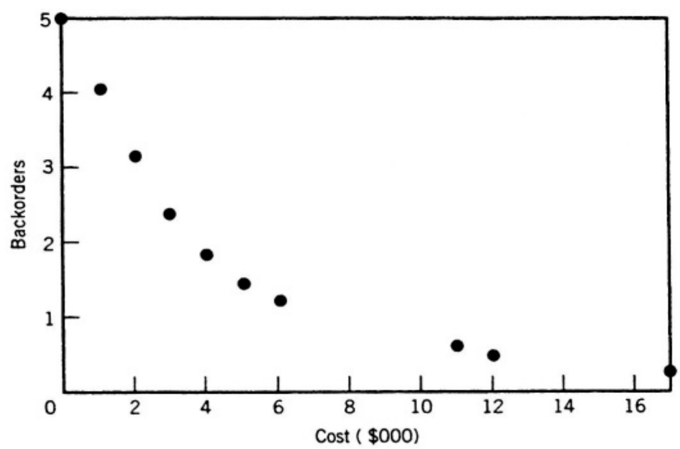


Figure 2-2. **Optimal system backorders vs. cost**

Figure 2-2 is not a continuous curve, but a set of points. The horizontal distance between adjoining points is determined by the cost of the item that is selected at that step of the iterative procedure. Sometimes the step size is 1, sometimes it is 5. How do we find the points in between?

We don't. This is a crucial distinction between solving a mathematics problem and a logistics problem. A mathematician would use a technique like dynamic programming, discussed in texts such as Hillier and Lieberman (1980), to find solutions for costs of 7, 8, 9, 10 and the other integral values (even the mathematician can't find solutions for non-integral values such as 7.3 and 9.8). We show below that all of our solutions are identical with those generated by the mathematician, but the mathematician will find solutions for intermediate points as well.

The logistician is a more practical character who realizes that we aren't really interested in two items, but many - perhaps thousands that comprise our system. Management may have said, "Don't spend more than a million dollars on spares", but that doesn't mean that an optimal solution costing $5000 less is non-responsive. Or if the logistician is told to achieve an availability of 95%, will he or she be chastised for determining the optimal spares that are expected to produce a 95.2% availability? In logistics applications we know that all of our data are estimates: demand rates, costs, lead times, repair times, probability of local repair capability (versus depot repair), scrap or condemnation rates, etc. We know that we will never hit the projected availability or cost precisely in the real world, regardless of the degree of mathematical sophistication that we employ.[1]

2.11 Convexity

Now we need to prove that marginal analysis produces optimal solutions. Note that in each of the two delta columns in Table 2-3, the numbers decrease steadily. Mathematically a function with this property is called *convex,* because that is the shape when viewed from the x axis.

A function $h(s)$, defined for discrete nonnegative s, is said to be convex if the first difference,

$$\Delta h(s) = h(s+1) - h(s)$$

is less than or equal to zero and the second difference

$$\Delta^2 h(s) = h(s+2) - 2h(s+1) + h(s)$$

is greater than or equal to zero. To prove that expected backorders is convex for any probability distribution, Pr{.}, and any argument s, we need only substitute Equation 2.7 into the definition:

$$\mathrm{EBO}(s+1) - \mathrm{EBO}(s) = \Pr\{DI = s+2\} + 2\Pr\{DI = s+3\} + \ldots$$

$$- \Pr\{DI = s+1\} - 2\Pr\{DI = s+2\} - 3\Pr\{DI = s+3\} - \ldots$$

$$\Delta\mathrm{EBO}(s) = - \Pr\{DI = s+1\} - \Pr\{DI = s+2\} - \Pr\{DI = s+3\} - \ldots \le 0 \quad (2.8)$$

$$\Delta^2\mathrm{EBO}(s) = \Pr\{DI = s+3\} + 2\Pr\{DI = s+4\} + \ldots$$

$$- 2\Pr\{DI = s+2\} - 4\Pr\{DI = s+3\} - 6\Pr\{DI = s+4\} - \ldots$$

$$+ \Pr\{DI = s+1\} + 2\Pr\{DI = s+2\} + 3\Pr\{DI = s+3\} + 4\Pr\{DI = s+4\} + \ldots$$

$$\Delta^2\mathrm{EBO}(s) = \Pr\{DI = s+1\} \ge 0 \quad (2.9)$$

[1] If an expensive item is bought at the last iteration, there may be a large overshoot of the target. This can be reduced by taking the stock levels on that item after the overshoot as minimum stock levels and rerunning the computation. The procedure can be repeated.

Since the expected backorder function is convex, the marginal analysis values $\{EBO(s - 1) - EBO(s)\}/c$, where c denotes the cost of the item, are non-increasing. The system backorders in Figure 2-2 are convex also; it is easy to show that the sum of convex functions is convex. (Problem 7).

Suppose that the backorder functions were not convex. The marginal analysis procedure of looking at the next improvement in backorders per dollar for each item could not guarantee an optimal solution. Consider Figure 2-3 and some arbitrary function $h(s)$. As drawn, $h(s + 1)$ is not a convex point, because it lies above the line segment Δ_{02}. When the stock level is s and we examine the improvement of moving to $s + 1$ using marginal analysis, it is only Δ_{01}. If we had looked two steps ahead and considered moving directly to $s + 2$, this item would have looked more attractive. We would have done the right thing, but only because we modified the marginal analysis.

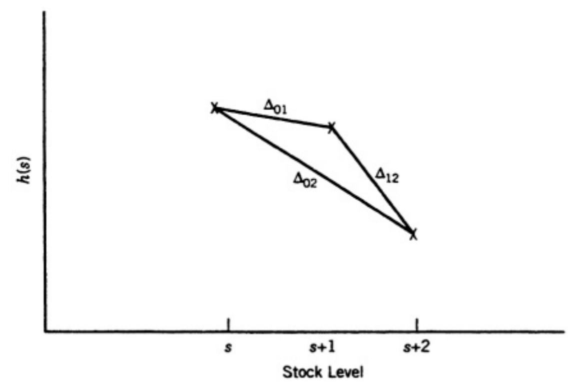


Figure 2-3. Nonconvex example

The problem with marginal analysis that looks only one step ahead is that there may be other items whose Δ are between Δ_{01} and Δ_{02}. Those item stock levels will be augmented prematurely. Of course, after we have moved from s to $s + 1$ on this nonconvex item, the next improvement to $s + 2$ will look very attractive because Δ_{12} is very large. At this point the overall solution has returned to optimality, though we cannot guarantee optimality at all intermediate points.

2.12 Mathematical Solution of Marginal Analysis

Now we proceed to the formal justification of the marginal analysis procedure. Its purpose is to show that a standard mathematical approach to this problem yields the same solution we discussed above. The mathematical statement[1] for two items is:

$$\min_{(s_1,s_2)} \mathrm{EBO}_1(s_1) + \mathrm{EBO}_2(s_2) \tag{2.10}$$

subject to:

$$c_1 s_1 + c_2 s_2 \leq C \tag{2.11}$$

for a series of total system cost targets, $C = 1, 2, 3, \ldots$

The standard procedure for minimizing a function subject to a constraint is to write the objective function from Expression 2.10 plus a Lagrange multiplier, λ, times the left-hand side of the constraint equation (Equation 2.11). Then if s_1 and s_2 were continuous variables, we would take the partial derivative with respect to s_1 and set it equal to zero; and with respect to s_2 and set it equal to zero. These are the necessary conditions for a minimum.

The analogue of the derivative for discrete variables is when the first difference satisfies:

$$\mathrm{EBO}_1(s_1) - \mathrm{EBO}_1(s_1 - 1) + \lambda c_1 \leq 0$$

$$\mathrm{EBO}_1(s_1 + 1) - \mathrm{EBO}_1(s_1) + \lambda c_1 > 0$$

or, equivalently

$$\mathrm{EBO}_1(s_1) - \mathrm{EBO}_1(s_1 - 1) \leq -\lambda c_1 < \mathrm{EBO}_1(s_1 + 1) - \mathrm{EBO}_1(s_1) \tag{2.12}$$

and a similar equation for s_2. But, these are precisely the quantities used in the marginal analysis. For any specified value of λ, the stock level s_1 that satisfies Equation 2.12 and s_2 for an analogous equation yield a point on the cost-backorder curve of Figure 2-2.

The conditions underlying Equation 2.12 are shown in Figure 2-4. For a fixed value of λ, the optimal stock level s_i^* for each item i has the property that the slope of the dashed line $\Delta\mathrm{EBO}_i(s_i^* - 1)/c_i$ is less than or equal to the slope $-\lambda$ and the slope of the dashed line $\Delta\mathrm{EBO}_i(s_i^*)/c_i$ is greater. (All slopes are negative). λ is a positive quantity and as it gets smaller, the stock levels increase. By contrast, the traditional item approach is to estimate λ as the

[1] There are two subscripts on each term; the first refers to the parameters of the underlying probability distribution for the item and the second refers to the stock level of the item.

annual holding cost rate divided by the annual cost of a backorder and find a single solution (see Problem 12).

These are the necessary conditions for a minimum. The sufficient conditions are satisfied by the convexity of the expected backorder functions (this is analogous to a continuous variable's second derivative being positive everywhere).

Note that the problem statement in Equations 2.10 and 2.11 could have been written for an arbitrary number of items, *I*. At each step of the marginal analysis there would be *I* quantities to compare, but the optimization procedure is still very efficient.

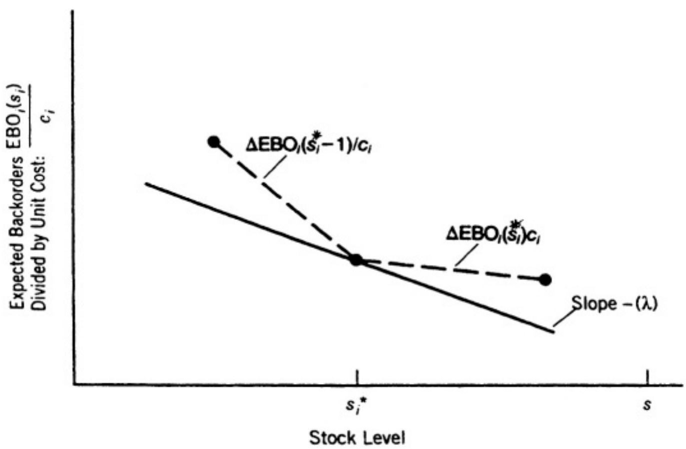


Figure 2-4. Optimality conditions: for any item i, s_i^* represents the optimal stock level. Δ is the slope. See text for details.

It is important to understand the graphical interpretation of Figure 2-2 and its relationship to the marginal analysis procedure. The absolute values of the slopes of line segments connecting adjacent solutions in Figure 2-2 are the delta values selected in the marginal analysis. These lines form the "convex hull" of the solution. The optimal solution for total cost of \$7000, \$8000, \$9000, or \$10,000 cannot be obtained with marginal analysis, and therefore must lie above the lines – i.e., not on the convex hull.

For example, in the problem discussed in Table 2-3 the optimal solution for a cost of \$7000 is easily found by trial and error to be 7 units of item 2. The system backorders for that solution are found from Table 2-3 to be 1.085. This is larger than the value of 1.067 obtained for a cost of \$7000 by interpolating on the line segment that would connect costs of \$6000 and \$11,000 in Figure 2-2.

In summary, the mathematician may want a solution for every integral value of cost. The logistician realizes this is unnecessary for his problem with thousands of items. Furthermore, the logistician knows that his marginal analysis procedure will find all of the "efficient" solutions on the convex hull.

2.13 Separability

One characteristic of the problem is key to the marginal analysis solution method. The objective function that we minimize is separable. In particular, the objective function is the sum of the backorders on all items. Thus when we take first differences on each decision variable, s_i, the optimal conditions involve only the backorders on item *i*. Had the objective function been more complicated, the optimal conditions for item *i* from Equation 2.12 might have involved other items and prevented us from using marginal analysis.

A function $h(s_1,s_2)$ is said to be additive separable if $h(s_1,s_2) = h_1(s_1) + h_2(s_2)$, where h_1 is a function of s_1 only and h_2 is a function of s_2 only. In the next section we consider an objective function that is not additive separable, but which can be converted into such a function by a logarithmic transformation.

2.14 Availability

Logisticians combine the word "availability" with three different adjectives -*inherent, achieved,* and *operational* - to measure different things. In order to understand availability in this book, it is critical to understand these usages.

$$\text{Inherent Availability} = \frac{100 \times \text{MTBF}}{\text{MTBF} + \text{MTTR}} \tag{2.13}$$

where MTBF is the mean time between failures and MTTR is the mean time to repair. Note this is a measure of hardware reliability and maintainability and has nothing to do with spares.

$$\text{Achieved Availability} = \frac{100 \times \text{MTBM}}{\text{MTBM} + \text{MCMT} + \text{MPMT}} \tag{2.14}$$

where MTBM is the mean time between maintenance, MCMT is the mean corrective maintenance time, and MPMT is the mean preventive maintenance time. The MTBM may be smaller than the MTBF, because it makes allowance for periods when the system will not be available due to preventive maintenance activities. While achieved availability is an improvement over inherent availability, it is a similar measure that relates to hardware considerations and excludes spares delays. Let us turn to the third type of availability, operational availability.

$$\text{Operational Availability} = \frac{100 \text{ x MTBM}}{\text{MTBM} + \text{MDT}} \tag{2.15}$$

where MDT = mean downtime due to spares, maintenance (corrective and preventive), and other delays resulting from maintenance action. Operational Availability is what this book is about. A system is operational if it is not down for either maintenance or supply. For calculational purposes, we will find it convenient to compute operational availability as the product of two availabilities, maintenance availability (identical to achieved availability in Equation 2.14) and supply availability where

$$\text{Maintenance Availability} = \frac{100 \text{ x MTBM}}{\text{MTBM} + \text{MCMT} + \text{MPMT}} \tag{2.16}$$

$$\text{Supply Availability} = \frac{100 \text{ x MTBM}}{\text{MTBM} + \text{MSD}} \tag{2.17}$$

where MSD is the mean supply delay. The mean delay time in Equation 2.15 is the sum of the delay times in Equations 2.16 and 2.17, i.e. MDT = MCMT + MPMT + MSD.

If either maintenance availability or supply availability is high, then the product is a good approximation to operational availability. For example, if maintenance availability and supply availability are each 95%, the product is 90.25% (after division by 100). A more precise calculation of operational availability use Equation 2.15. This requires an estimate of the components of MDT from Equations 2.16 and 2.17. Thus, using the maintenance availability of 95% in Equation 2.16, we find (MCMT + MPMT)/MTBM as .0526. Similarly, we use the supply availability of 95% in Equation 2.17 to estimate MSD/MTBM as .0526. Then the operational availability from Equation 2.15 is 1/(1 + .1052) = 90.48%. The product of Equations 2.16 and 2.17 understates operational availability by a larger amount as the availabilities decrease. Thus, if maintenance availability and supply availability are each 90%, the product is 81% (after division by 100), whereas the more precise calculation yields 81.8%. The more precise computation is easy to perform and is recommended if at least one or the availabilities is low.

We break operational availability into these two availability components, because it simplifies computation. Once the maintenance manning, test equipment, and preventive maintenance policy have been defined, the maintenance availability can be calculated. It is a single number that depends on the mean time between maintenance, MTBM, but is independent of the stockage policy. The calculation of maintenance availability is addressed in Section 5.16 for the case of continuous resupply and in Section 7.3 for periodic resupply.

Supply availability is independent of the maintenance policy, but it is not a single number. We compute supply availability as a function of the stockage policy, and it is this optimal availability-cost curve that we compute below. Whenever we use the term *availability* in this book without a qualifying adjective, it is supply availability to which we refer.

In Chapter 1 we defined *cannibalization* to be the consolidation of any line-replaceable unit (LRU) "holes" onto the smallest number of aircraft. Now we show that the minimization of total backorders is almost equivalent to maximization of availability when cannibalization is not practiced. Note that logisticians use the term "availability" to denote an expected value (over the probability distribution of demand), so we will not use the term "expected availability" here[1]. Availability, A, the expected percent of the aircraft fleet that is not down for any spare is given by the following product:

$$A = 100\prod_{i=1}^{I}\{1 - \mathrm{EBO}_i(s_i)/(NZ_i)\}^{Z_i} \qquad (2.18)$$

with the understanding that $A = 0$ if $\mathrm{EBO}_i(s_i) > NZ_i$ for any item i. Z_i is the number of occurrences on an aircraft of the ith LRU (quantity per aircraft) and N is the number of aircraft. The logic is that there are NZ_i locations of LRU i in the fleet of aircraft, the probability of a hole in any of these locations is $\mathrm{EBO}_i(s_i)/(NZ_i)$ (the probability cannot exceed one). An aircraft will be available only if there is no hole for any of the Z_i occurrences of LRU i (which accounts for the exponent), or for any other LRU (which accounts for the product over i).

Note that Equation 2.18 was computed on the assumption of independence of failures across aircraft and no cannibalization. When the

[1] In Chapter 6 we consider periodic resupply. Availability is still an expected value over the demand probabilities, but availability decreases over the cycle between resupply missions. We use the term "average availability" in that section to denote an expectation over time.

fleet size N is greater than one, the availability is the expected percentage of the fleet that is operational; we can also think of the availability as the expected percentage of time that any aircraft in the fleet is operational. Taking logarithms in Equation 2.18, and remembering that the logarithm of a product equals the sum of the logarithms yields:

$$\log(A/100) = \sum_{i=1}^{I} Z_i \log\{1 - \mathrm{EBO}_i(s_i)/(NZ_i)\} \cong -\sum_{i=1}^{I} \mathrm{EBO}_i(s_i)/N \tag{2.19}$$

where the last approximation comes from the power series expansion of $\log(1 - a) = - a - .5a^2$ ..., ignoring terms in a^2 and higher powers of a since a will be small in the cases of interest.[1] For example, if a is less than .1, the error of dropping all the terms in a^2 and higher is only about .005. Usually we are interested in availabilities of at least 80%, which implies that the value of a will be much smaller than .1; a value of a = .1 on a single item brings the availability down to 90% even before the impact of other items is taken into account.

Thus the logarithm of availability is a convex, additive separable function of the item backorder functions. Since a function and its logarithm achieve their maximum at the same point, we can maximize the logarithm of availability in Equation 2.19 and this is equivalent to maximizing availability itself. Furthermore, Equation 2.19 shows that this maximization is accomplished by minimizing the sum of item backorders. Figure 2-5 showsthe optimal availability vs. cost curve, which is obtained from the optimal backorder vs. cost curve of Figure 2-2 by substituting the expected backorders into Equation (2.18), where we have assumed that N = 10 and each $Z_i = 1$. For example, at a cost of 0 the availability in Figure 2-5 is just the product 100(1 – 1/10)(1 - 4/10) = .54.

It is important to emphasize that each discrete point on the optimal availability vs. cost curve in Figure 2-5 is the maximum availability for the specified cost, and equivalently the minimum cost for that availability.

[1] The power series expansion is valid for a<1 when the logarithm is taken to the base e = 2.718. The logarithm of a number is the exponent to which the logarithmic base is raised. Thus to find the logarithm of 100 to the base e, we need the value a satisfying $e^a = 100$. It is clear that any number B larger than 100 must have a logarithm, b, larger than a, showing that a number and its logarithm achieve their maxima at the same point. Furthermore, since $100B = e^a e^b = e^{a+b}$, or log (100B) = a + b, the logarithm of a product is the sum of the logarithms.

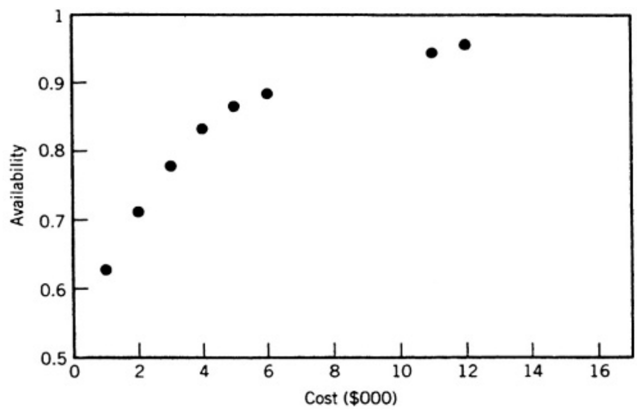


Figure **2-5. Optimal system availability vs. cost**

2.15 Summary

We have shown how Palm's theorem can be used to calculate the steady-state probability distribution for the number of units of an item due in from repair. Note that this distribution does not depend on the probability distribution for the time to repair an item, but only the mean of the distribution. Knowledge of the steady-state distribution for the number of units in repair enables us to estimate the probability distribution for stock on hand and that for backorders. It allows us to compute the fill rate also.

It was shown that minimization of total backorders is approximately equivalent to maximization of availability. Furthermore, due to the convexity and separability of the backorder function, we are able to use marginal analysis to generate the optimal availability vs. cost curve.

The marginal analysis procedure finds all of the efficient solutions on the "convex hull". However, there are interior points for intermediate cost values which can only be obtained by combinatorial procedures such as dynamic programming. Though these solutions may be of interest to the mathematician, they have limited significance to the logistician whose systems may comprise thousands of items.

We showed that operational availability can be computed from maintenance availability and supply availability. Maintenance availability is a single number that depends on the maintenance resources, but is independent of the supply policy. Supply availability is independent of maintenance resources, but it is not a single number. It is a function of the supply policy, and it is this optimal availability vs. cost relationship that we focus on in this book.

Finally, it is important to remember that each discrete point on the optimal availability vs. cost curve in Figure 2-5 is the maximum availability for the specified cost, and equivalently the minimum cost for that availability.

2.16 Problems

1. Use the Poisson definition from Equation 2.4 and the formulas for mean and variance in Equations 2.1 and 2.2 to compute their values for the Poisson distribution.

2. Typically when probabilities are needed to evaluate backorders or fill rates, we need a series such as $\Pr\{X = x\}$, $\Pr\{X = x + 1\}$, etc. Instead of calculating each probability from the formula in Equation (2.4), find a recursion formula for the Poisson which for any x expresses $\Pr\{X = x + 1\}$ in terms of $\Pr\{X = x\}$ and some multiplicative factors. Use the recursion to show that the Poisson is unimodal, and the relationship of the mode to the mean.

3. Recalling that EFR(s) is the expected fill rate as in Equation 2.6, develop the recursion formula for EFR(s + 1) in terms of EFR(s) and multiplicative terms; and for EBO(s + 1) in terms of EBO(s) and multiplicative terms. Use the latter recursion to compute the first four values in Table 2-1 for an average pipeline of 1.

4. Show that fill rate is not convex for small values of s when the average pipeline exceeds one, by examining EFR(s - 1) - EFR(s). For the Poisson distribution, find the limiting value at which s becomes convex as a function of the average pipeline value.

5. Use trial and error to find expected backorder solutions for costs of \$8000, \$9000, and \$10,000 in Figure 2-2, and show that these are larger than the values on the line segment of the convex hull in Figure 2-2 at the corresponding costs.

6. Suppose that there has been a price decrease to \$1000 for item 1 in Table 2-1. Show what changes this produces in Table 2-3 and then obtain a new cost vs. backorder curve to replace Figure 2-2. Now take these optimal system backorder values and plot them on a copy of the original Figure 2-2 using the original cost of \$5000 for item 1 (this is equivalent to ignoring unit cost in the marginal analysis, since both items had a cost of \$1000). Show that when unit cost is ignored in the marginal analysis, the new "solution" is never better and sometimes worse than the original.

7. Show that the sum of two convex functions must be convex. (*Hint*: apply the definition of a convex function)

8. Use the table for expected backorders with a mean of 1 from Table 2-3 and the expected backorders with a mean of 10 given below to verify the

optimality of the no-cannibalization policy in Table 1.4 and the availability. Also evaluate under the assumption of no cannibalization the availability of the other two stockage policies.

Stock level	0	1	2	3	4	5	6	7
Backorders	10.000	9.000	8.001	7.003	6.014	5.043	4.110	3.240
Stock level	8	9	10	11	12	13	14	15
Backorders	2.460	1.793	1.251	0.834	.0531	0.323	0.187	0.104

9. Instead of requiring that each demand be met at the time it is placed to count as a fill, define a "delayed fill" as a demand that is met within a specified time interval t or less. Determine if Palm's theorem can be modified to accept this change, and the effect on the fill rate and backorder Equations 2.6 and 2.7. Is this result still independent of the shape of the repair time distribution? Does this affect the infinite-channel queueing assumption? As t gets larger what does the delayed fill rate measure begin to resemble? The answers are given in Feeney and Sherbrooke (1966) and Section 9.8. See also Chapter 6 where models based on periodic resupply are considered.

10. The purpose of this problem is to illustrate that unit cost should affect the stockage policy even when demand is not probabilistic. Suppose that there are two items, each of which has one demand at the beginning of every week and each requires four weeks to repair. The unit costs of the two items are \$10 and \$100 respectively.

Suppose there are 20 aircraft, each having only one unit of each item. If we had a budget of \$440, we could have 100% availability by filling the average pipelines. Assuming that cannibalization is not allowed, develop the optimal cost vs. availability curve for spares budgets between 0 and \$440.

11. Develop the optimal cost vs. availability curve up to an availability of 95% for Problem 10 (above) under the same assumptions, except that demand is Poisson.

12. Schultz (1990) has derived the conditions under which the $(s - 1, s)$ inventory policy is optimal. Suppose the demand is Poisson with average demand mT over the lead time, the annual holding cost per item is ιc (the annual holding cost rate times the cost of the item), and the annual cost of a backorder is π. Then one must solve the following inequality for the smallest value of s satisfying:

$$\sum_{i=0}^{s} \frac{(mT)^i}{i!} + \frac{(mT)^{s+1}}{2(s+1)!} \geq \frac{\pi e^{mT}}{\pi + \iota c} \tag{2.20}$$

Then the solution s^* must be plugged into a second inequality, where the $(s-1, s)$ policy will be optimal for any order cost, Ω, that satisfies:

$$\Omega \leq \frac{\pi + \iota c}{m} \left| e^{-mT} \sum_{i=0}^{s^*} \frac{(mT)^i}{i!} - \frac{\pi}{\pi + \iota c} \right| \tag{2.21}$$

For example, if $mT = 1$, $\pi = 700$, and $\iota c = 300$ then $s^* = 1$ from the first inequality (Equation 2.20). When $m = 1$, we find from the second inequality that the $(s - 1, s)$ policy is optimal if $\Omega \leq 35.8$.

What is the largest admissible value of Ω if $\pi = 7000$ and $\iota c = 3000$? What is the value of Ω if $\pi = 7000$, $\iota c = 3000$, $mT = 10$ and $m = 10$? Contrast these probabilistic results to the deterministic demand case of Problem 1 in Chapter 1. In Chapter 5 we develop approximate methods to generalize the $(s - 1, s)$ assumption.

Chapter 3

METRIC: A MULTI-ECHELON MODEL

Science is a first-rate piece of furniture for a man's upper chamber, if he has sense on the ground floor.

– Oliver Wendell Holmes

3.1 Chapter Overview

In 1966, while at the RAND Corporation, the author developed the Multi-Echelon Technique for Recoverable Item Control (METRIC) for the Air Force (Sherbrooke 1968). It extends the single-site model discussed in Chapter 2. First-indenture items (i.e. parts that are installed directly on aircraft) are usually repairable (or recoverable), and they tend to be expensive, with low demand at any particular base. Thus, there is little batching of repair at the base level or batching of resupply requests on depot for multiple units of a given item. The one-for-one repair at base or resupply from depot simplifies the mathematics of the base-depot joint optimization problem. It turns out that the inventory problem can be addressed with results from the queueing theory literature.

The METRIC theory is the basis for a large number of multi-echelon models used by the several military services. In Chapter 4 we show how to model the multi-indenture problem (where the repair of first-indenture line replaceable units (LRUs) can generate demands for lower indenture parts), and how this can be made more accurate by taking into account the dependencies between indentures using VARI-METRIC theory. Then in Chapter 5 the VARI-METRIC theory is applied to the combined multi-echelon, multi-indenture problem. The mathematical details underlying METRIC are kept to a minimum in this chapter, because these are provided in Chapter 5, when we derive the more accurate VARI-METRIC theory.

There are some applications where a repair/resupply request can not be initiated when a failure occurs. We consider the case of resupply that can take place only at certain points in time in Chapter 6 when we consider Space Station *Freedom.*

3.2 METRIC Model Assumptions

The METRIC theory calculates for every item on a system the optimal stock level at each of several bases, which may be different in terms of item demand rates and other characteristics, and the supporting depot. The objective function is the sum of backorders across all bases, and we know from the previous chapter that minimization of base backorders is equivalent to maximization of availability when there is no cannibalization. As in Chapter 2, we use the system approach where management is presented with the optimal availability-cost curve.

Before developing the mathematics of METRIC, we want to list the key assumptions. These assumptions will be violated occasionally in practice, but they should be true most of the time or a different model should be used.

1. *The decision as to whether a base repairs an item does not depend on stock levels or workload.* The assumption is that some fraction of repairs on each item are within the base capability to repair, because they have the test equipment, personnel, and other resources. Whenever the base has the capability, repair is accomplished there *regardless of the maintenance workload.* If the necessary spare parts are not available at base, the base is supposed to requisition the parts from the depot.

2. *The base is resupplied from the depot, not by lateral supply from another base.* This appears to be appropriate for setting stock levels, because the number of lateral shipments is typically small and they are apt to induce special expediting costs. When lateral supply is ignored, transportation costs

between depot and bases are not needed, because the total transportation cost is not a function of the stockage policy. In Appendix B we develop procedures to evaluate the reduction in backorders that occurs when lateral supply is allowed. This is employed in the VMetric model of Appendix E.

3. *The (s - 1, s) inventory policy is appropriate for every item at every echelon.* The demand rates are sufficiently low and the costs are sufficiently high that the economic order quantity of Equation 1.1 is close to one. This means that units of an item are not batched for repair, and that any scrapped units are reprocured on a one-for-one basis.

Naturally some deviation from these assumptions is allowable. In the case of repair at the depot, there may be some setup time to install the adapter on automatic test equipment for the item to be fixed. If the item demand at depot is moderately high, maintenance management will want to repair several units of the item in a batch. In such a case the average repair time used in the model should include the average waiting time before depot repair is initiated.

Typically the condemnation or scrap rate for first-indenture items is quite low. The condemnation rate must be considered for procurement purposes, but we assume in this chapter that an (s - 1, s) policy is used at depot whereby one unit is procured to replace each condemnation as it occurs. In Chapter 5 we relax the assumption of the (s - 1, s) inventory policy, and this is particularly important at depot for lower cost, lower-indenture items with appreciable condemnation rates.

4. *Optimal steady-state stock levels are determined.* The assumption is that over some period of time in the future the number of aircraft or flying hours will remain fairly consistent. If the factors that contribute to demand are changing, we may want to compute for the "peak program", but the stock levels that emerge are steady-state levels. As discussed in Chapter 4, this does not mean that the demand rates for each item must be constant.

The original METRIC paper discussed compound Poisson demand, a generalization of the Poisson where demands occur in clusters. The number of demands in a cluster is given by a probability distribution such as the geometric; the probability distribution of time between clusters is exponential. The Poisson can be thought of as the special case of compound Poisson where the cluster size is always one.

We do not use compound Poisson processes in this book, because their "physics" is inconsistent with what we observe in applications. This is described in more detail in the next chapter, where we note that the probability distribution for the number of demands in time T, known as the *state probabilities*,[1] has an observed variance-to-mean ratio that tends to

[1] The term *state probabilities* is used because the states have discrete values corresponding to 0, 1, 2, . . . demands

increase with T. By contrast, the variance-to-mean ratio is constant for stationary Poisson and compound Poisson processes.

3.3 METRIC Theory

The theory will be developed in two steps: 1) for a single first-indenture item we develop the theory for optimal allocation of stock levels between the several bases and supporting depot, i.e we show how to construct an optimal cost-backorder curve for a single item; 2) we combine all items on a system using marginal analysis in a manner precisely analogous to the methods used in Chapter 2. This optimal system cost-backorder curve can be converted into an optimal system availability-cost curve in the same manner also.

We define the following variables for a single item:

m_j = average annual demand at base j

T_j = average repair time (in years) at base j

μ_j = average pipeline at base j

r_j = probability of repair at base j

O_j = average order and ship time from depot to base j

and we use the convention that positive subscripts refer to bases, a 0 subscript refers to the depot. The pipeline at base consists of due ins from both repair and resupply. At the depot the the due ins are assumed to be in repair; in Chapter 5 this assumption is generalized to allow for condemnations and procurement.

As noted earlier, the order-and-ship time is defined to be the time from placing a resupply request at the base until the item is received from the depot, *if the depot had stock on hand when the resupply request was received.*

Of course, the depot does not always have stock on hand when a resupply request is received. METRIC computes this depot delay in resupplying bases, which depends on the depot stock level. Then we can compute the backorders at each base which depend on the resupply delay from depot and the base stock level.

First we must calculate the average demand on the depot, and this is just the fraction of demand that is not repairable at each base, summed over the J bases:

$$m_0 = \sum_{j=1}^{J} m_j (1 - r_j) \tag{3.1}$$

We are assuming that the demand is given by a Poisson process, and using the fact that a sum of Poisson processes is a Poisson process. The average depot pipeline is $m_0 T_0$, and the expected backorders at depot[1], $\mathrm{EBO}(s_0 | m_0 T_0)$, are the expected number of base resupply requests that are outstanding at the depot at a random point in time. Another way of looking at this quantity is that it is the average delay added daily to resupply requests, resulting from the fact that the depot does not always have stock on the shelf.

Now we can compute the average pipeline for each demand at base j as:

$$\mu_j = m_j (r_j T_j + (1 - r_j)\{O_j + \mathrm{EBO}[s_0 | m_0 T_0] / m_0\}) \qquad j>1 \tag{3.2}$$

If the item is always base repairable, the average pipeline for base j is $m_j T_j$, the average number of units in base repair; otherwise, the average pipeline is $m_j O_j$, because of the order-and-ship time, plus the expected backorders for base j at the depot due to the fact that the depot does not always have a unit on the shelf. Note that when the depot stock level is very large, the depot backorders are zero; when the depot stock level is zero, the average delay time added to each demand is T_0, because failed unit must be received from a base and repaired. (The depot repair time is assumed to include a time allowance for shipment from base to depot of the failed unit).

Before leaving METRIC theory, we should note that there are a number of variables that are not needed in the optimization. Order costs and holding costs are not needed, because one-for-one replenishment is assumed and this defines the number of orders and the average stock on hand. Repair costs are not needed, because we assume that if an item can be repaired, its repair cost is less than its purchase cost.

3.4 Numerical Example

We illustrate this theory with an example. Suppose that are five identical bases with Poisson demand and the following parameter values for all $j \geq 1$:

[1] Note that this expression for expected backorders is an alternative to the form used in Equation 2.10 where a subscript on expected backorders was used to indicate the item; here the vertical line means that information to the right is the given information, namely the mean of the Poisson distribution.

m_j = 23.2 demands/year

T_j = .01 years

r_j = .2

O_j = .01 years

T_0 = .02531 years

Of course, each data element for a base could differ, but the computation of the example is simplified when they are identical. From Equation 3.1 we determine that $m_0 = 92.8$ and $m_0T_0 = 2.349$. We start with a depot stock level of 0 and compute μ_j for any base from Equation 3.2:

$$\mu_j = 23.2\ \{(.2)(.01) + (.8)[.01 + 2.349/92.8] = .7017 \qquad (3.3)$$

From this we can calculate the expected backorders, EBO(s), and the first differences, EBO(s - 1) - EBO(s), in Table 3-1, which is precisely analogous to Table 2.3. A backorder table for stock levels of 0-2 at any base is shown in Table 3-1.

Table 3-1. Expected backorders at any Base (Depot Stock Level = 0)

s	EBO(s)	EBO(s - 1) - EBO(s)
0	.7017	
1	.1975	.5042
2	.0411	.1564

The only change from Table 2.3 is that we do not need to divide the first differences by unit cost to apply marginal analysis, because we are considering only one item whose cost does not vary by base. In this particular example where all the bases are identical, the marginal analysis is even easier to apply because of symmetry. After marginal analysis we find the first eight points to be as shown in Table 3-2.

The expected backorders with no stock is just the sum of the backorders at each base with no stock (5 x .7017 = 3.5087). Then the stock level at each base successively is increased to one, then to two, etc. For example, the sixth unit of stock optimally allocated goes to base 1 and results in total backorders of 0.8309. We know this procedure will produce an optimal solution, because it was shown in Chapter 2 that the backorder function is always convex.

Table 3-2. Optimal Expected Backorders for Depot Stock Level = 0

	Total Stock at Bases Optimally Allocated								
Depot St. Level	0	1	2	3	4	5	6	7	8
0	3.5087	3.0044	2.5022	1.9959	1.4916	.9873	.8309	.6745	.5181
Optimal Base		1	2	3	4	5	1	2	3

In this manner we compute the expected backorders at bases when stock is optimally allocated to them, conditional on the fact that the depot stock level was assumed to be zero. We keep track also of the optimal policy that led to these backorders by noting the base to which each successive allocation is made.

Now we must repeat the procedure for a depot stock level of 1, which means recalculating μ_j from Equation 3.2. The only change in Equation3.3 is that EBO(1|2.349)] = 1.444 replaces EBO(0|2.349)] = 2.349 as the last numerator. We compute the new pipelines, μ_j, corresponding to this depot stock level of one and an expected backorder table similar to Table 3-1, apply marginal analysis and obtain the next row in Table 3-2. Continuing this procedure for larger values of depot stock level yields Table 3-3.

Our objective is to compute the optimal allocation of stock across bases and depot. Since each entry in Table 3-3 is the expected backorders for a specified level of depot stock and a specified level of stock *optimally allocated to bases,* we need only pick the minimum backorders on each diagonal running from upper right to lower left. Each value on a given diagonal is for the same amount of total stock, divided between the depot and the bases (optimally). For example, with a total stock of 2 the optimal backorders for depot stock levels of 0, 1, and 2 are 2.5002, 2.1983, and 1.9240 respectively in Table 3-3. Since 1.9240 is the smallest value, it appears in Table 3-4.

Table 3-4. Optimal Expected Backorders

Total Stock	Depot Stock	Stock at Bases	Total Backorders	Backorder Reduction	Convexity	Marginal Analysis[a]
0	0	0	3.5087		*	*
1	1	0	2.6043	0.9044	*	*
2	2	0	1.9240	0.6803	*	*
3	3	0	1.5072	0.4168	*	*
4	3	1	1.2469	0.2602		
5	2	3	0.9658	0.2811		
6	1	5	0.5743	0.3915	*	
7	2	5	0.3269	0.2474	*	
8	3	5	0.2060	0.1209	*	*

[a]See text for explanation of this column (Section 3.5)

Table 3-3. Optimal Expected Backorders for any Depot Stock Level

Table 3.3.

Depot Stock	Total Stock at Bases, Optimally Allocated								
Level	0	1	2	3	4	5	6	7	8
0	3.5087	3.0044	2.5002	1.9959	1.4916	0.9873	0.8309	0.6745	0.5181
1	2.6043	2.1983	1.7923	1.3863	0.9803	0.5743	0.4777	0.3811	
2	1.9240	1.6046	1.2852	0.9658	0.6464	0.3269	0.2694		
3	1.5072	1.2469	0.9867	0.7264	0.4662	0.2060			
4	1.2965	1.0681	0.8397	0.6113	0.3829				
5	1.2070	0.9925	0.7780	0.5636					
6	1.1743	0.9650	0.7557						
7	1.1639	0.9562							
8	1.1610								

The optimal allocations of stock for each policy in Table 3-4 were determined during the marginal analysis. Since the optimal allocation to bases of a given total stock at bases is obvious from symmetry, we have not shown that breakdown in Table 3-4. We show the total backorders at all bases for each optimal policy and the reduction in total backorders at each step.

3.5 Convexification

Note that the backorder reductions are not monotonically decreasing, which means that the backorder function is not convex at those points. We have indicated this by omitting the corresponding asterisks in the convexity column. Why does this non-convexity occur and what do we have to do because of it?

Each row of Table 3-3 is convex, because the marginal analysis procedure is identical with Chapter 2. But as we move from one level of depot stock to another in the optimal solutions, there is no assurance of convexity. In fact that is why we chose an example with several identical bases, because it is more likely to occur in such a case.

The nonconvexity is easily identified by scanning the Backorder Reduction column in Table 3-4. These points are easily dealt with by excluding them as potential solutions. The points that remain are similar to Figure 2.2 in that the line connecting adjoining solutions has a slope that always flattens out as the total stock optimally allocated increases.

The reason we need to check for non-convexity and discard such interior points is that the marginal analysis procedure for combining items will be misled otherwise. As exemplified in Figure 2.3, when we reach a non-convex point the slope will be flatter than if we look at the next convex point. Thus, the backorder reduction per dollar for that item will be understated. By dropping the interior points, the marginal analysis will jump to the next convex point at the correct time (buying at least two more units of stock of the item because of the eliminated interior point or points). It is easy to check for these nonconvex points for total stock of 4 and 5 units in Table 3-4 and eliminate them in the computer algorithms.

Once we have eliminated the nonconvex points for each item, as in Table 3-4, we can use marginal analysis again as in Chapter 2 to optimize across items. Note that when we combine across items, the backorder reductions (first differences) must be divided by unit cost.

The last column of Table 3-4 concerns solutions obtained by marginal analysis used in a different manner. Suppose that instead of generating a complete table of solutions such as Table 3-3, we move sequentially from one solution to the next by either allocating a unit of stock to the depot or one to each base (since the bases are identical). Note that after the solution with 3 at depot and none at bases, such a marginal analysis procedure would next find the solution with 3 at depot and 1 at each base, missing the intermediate solutions.

This simple marginal analysis solution produces efficient solutions here, but a less dense set then generated by the exhaustive search procedure (the pattern repeats at larger values of total stock). It is possible to construct examples where some of the solutions are not optimal. However, one nice feature of the simplified procedure is that by construction the solutions must be convex.

The phenomenon observed concerning optimal stockage policies in Table 3-4 is referred to as *flushout.* After three units of stock have been allocated optimally to depot and more stock is available for allocation, it becomes optimal to take some of that depot stock and flush it out to the bases. (In this case the optimal depot stock drops to one when there are a total of six units to be allocated). Another reason for using an example with identical bases was to illustrate the flushout phenomenon which is more common in that situation.

3.6 Summary of the METRIC Optimization Procedure

Now to summarize the optimization procedure:

1. Start with a depot stock level of zero.

2. Compute the average resupply delay at the depot due to the fact that the depot does not always have stock on the shelf, and the average response time to each base from Equation 3.2.

3. Calculate the expected backorders for each level of base stock. Repeat for each base. (Table 3-1)

4. Use marginal analysis to combine the base backorder functions and obtain the minimum backorders for each number of units at bases. (Row in Table 3-3 corresponding to the depot stock level).

5. If the level of depot stock is large enough, go to step 6; otherwise, increase the depot stock level by one and go to step 2.

6. Find the minimum value on each diagonal representing the same number of units of stock. (Table 3-4) Drop any non-convex points.

7. Repeats steps 1-6 for each item.

8. Use marginal analysis to combine the item solutions (Table 3-4), where the first differences must be divided by the item costs.

3.7 Availability

We have developed procedures for minimizing backorders across bases and items. Now the relationship to availability, A, must be demonstrated. In Chapter 2 we maximized the availability at a single base; here we maximize the availability across all bases. First we modify Equation 2.18 for the availability at a single base by appending a subscript j to indicate the specific base. Then the overall system availability, A, is just the percentage of aircraft that are operational at all bases.

$$A = \sum_{j=1}^{J} A_j N_j / \sum_{j=1}^{J} N_j \tag{3.4}$$

Unfortunately, this objective function is the sum across bases of several product functions, one for each base. Thus we cannot take logarithms and expect the calculation to break into separable pieces. However, as a practical matter we can get a good approximation to the optimal allocation sequence by finding the maximum across bases and items of the following: the increase in availability times the number of aircraft divided by the cost of the item. Thus we can use marginal analysis again.

The availability at each base that results from this procedure will differ somewhat from one base to another. If it is desired to have nearly the same availability at each base, it is possible to achieve this by multiplying the availability at the bases that were below the average by weights that are larger than one. Or more generally it is possible to approach any desired set of availabilities that may differ by base. Of course, it will be necessary to try several sets of weights to achieve a particular set of base availability goals, and because the number of items and stock levels are discrete, the goals cannot be met precisely.[1]

[1] The procedure described above produces an optimal solution. The weights are Lagrange multipliers whose magnitudes represent the difficulty in meeting the constraints at the corresponding bases. However, adjusting the set of weights and rerunning the optimization can be tedious.

In the VMetric model, described in Appendix E, it is possible to specify availability targets by base and/or by aircraft type, when there are multiple aircraft types. This is due to an ingenious procedure developed by John Millhouse, TFD Group, which automatically

3.8 Summary

In this chapter we have extended the theory from Chapter 2 to cover the multi-echelon problem. The same formulas for expected backorders are used, based on Palm's theorem that says we do not need to know the shape of the repair/resupply distribution.

Since the sum of backorders at bases is an additive separable function of the items, we can focus on one item at a time. For an item, the formulas are applied first to the depot where the average delay to each base is computed as a function of the depot stock level. Marginal analysis is used to determine the optimal allocation of stock levels to bases for each specific level of depot stock. Then the minimum value of total base backorders is selected from each diagonal representing the alternative allocations of a given number of units of stock.

The resulting backorder function may not be convex, but it can be convexified by dropping solutions that are not convex. The analysis is repeated for each item, and then marginal analysis is applied a second time to find the optimal allocation of investment across items.

3.9 Problems

1. Calculate Table 3-1 for a depot stock level of 1, use marginal analysis, and show that the second line of Table 3-3 is obtained.

2. Outline the logic for a three-echelon calculation. Consider the number of calculations if all bases are identical and the number of bases supported by each intermediate site is the same. Compare this with a case where both assumptions are relaxed.

3. Instead of using an availability or cost target for the optimization, it is sometimes possible to determine the optimal point on the cost-availability curve. For example, if the end-item is a commercial aircraft, it may be possible to estimate the cost of the aircraft being down (e.g. the rental cost of a replacement). Suppose that the next optimal allocation of investment is to buy an item costing c dollars for a particular base or depot, where the annual cost of that investment is the holding cost ι times the item cost. Let the corresponding increase in availability of end-items be denoted ΔA and the number of end items be denoted by N.

Write an expression for the marginal annual cost saving in aircraft down by buying the next spare and equate it to the marginal annual cost of buying

---

reduces the desirability to the marginal analysis of a base or aircraft type when it approaches its target. This procedure, though elegant and easy to use, can be improved to an optimal constrained solution by iterating on base essentialities as illustrated after Figure E-6.

the spare. (A fundamental theorem of economics is that profit is maximized when marginal cost equals marginal revenue). Solve for the increase in availability per dollar (the slope of the cost-availability curve) at this optimal point. Assume that $\iota = .2$/year, there are $N = 100$ end-items, and the imputed cost of an aircraft down for a year is \$200,000. Show that the optimal point on the availability-cost curve is where the slope equals .00000001 = 1.0E-08.

It should be noted that aircraft may be down due to maintenance as well as supply. The fraction of the time down due to maintenance is described in Section 5.15, but for simplicity assume that the downtime cost D used above is for downtime due to supply.

Is this approach still valid when the end-items are military systems? What is the role of sunk cost?

Note that the annual cost of ιc, used above, assumes that the aircraft will be in service for many years. However, if the system has a life of 15 years or less, it is necessary to use a larger annual cost which amortizes the investment. Consider how an amortization formula might be used.

Chapter 4

DEMAND PROCESSES AND DEMAND PREDICTION

Take calculated risks. That is quite different from being rash

– Patton

4.1 Chapter Overview

When the George AFB field test, described in Chapter 1, was conducted, several people asked us why an inventory model was needed at all. They suggested that to find out what is needed all one has to do is ask an expert, the supply sergeant. There was one particular sergeant who had kept a little black book after a thirty day deployment exercise. In his words, "We took a lot of stuff we didn't need, and we needed a lot of stuff we didn't take. But I know exactly what we need next time, and that's all we're taking."

The sergeant's data could certainly be useful, but his understanding of the problem was flawed because he did not recognize the importance of variability. If items with demand during the previous exercise are the only ones taken on the next exercise, it is certain that there will be demands for other items.

Expert judgment may provide many important inputs to a stockage model. But experts tend to be very poor at processing probabilistic information intuitively, or even recognizing whether a stream of data is from a random process. When it comes to making decisions about the next item to buy and where to put it in a multi-echelon, multi-indenture problem where demand is probabilistic and the mean may be changing, a model is indispensable.

In the early 1960's there were few inventory modeling projects underway and a lot of demand prediction studies. Now the situation seems to have reversed, with very little attention to demand prediction. This is partially because of the success we have had in building ever more complicated inventory models and implementing them on personal computers. But it is important to spend a portion of our overall effort in demand prediction studies. After all, the best inventory model in the world isn't going to be much help if the demand estimates are garbage.

In the simplest case of random demands with a constant mean, an exponential probability distribution for the time between demands leads to Poisson (state) probabilities for the number of demands during any specified time period, T. This is called a Poisson process.

The Poisson distribution can be generalized to the negative binomial whose variance exceeds the mean. The negative binomial has two parameters, allowing us to fit a mean and variance separately to observed data. This is useful for two reasons: (1) the number of units of an item in the pipeline in Chapter 3 was treated as if it were a Poisson variable, but the observed variance usually exceeds the mean, often substantially; (2) when the demand rate changes or drifts with time leading to a Poisson process with non-stationary increments, the variance will exceed the mean.

The multi-indenture problem is used to illustrate the improvement in accuracy when variances in pipelines are taken into account. This is the foundation of VARI-METRIC theory, the subject of Chapter 5. (The VARI prefix is used because the calculation includes pipeline variances).

From earlier demand prediction studies we know that over very short periods, demand for most items tends to be Poisson with a constant mean, and that the variance-to-mean ratio for a particular item tends to increase as the period of observation lengthens. This implies that the demand process for most items is Poisson with non-stationary increments where the mean is not constant but drifts with time. The probability distribution for the number of demands in any time period can be modeled with the negative binomial distribution.

In order to estimate this drift, we performed several demand prediction experiments where the mean and variance for each item were estimated. Prior to that discussion, it is necessary to describe the "objective" Bayes technique that was tested, along with an important, related procedure known as James-Stein estimation. The demand prediction experiments provide some empirical evidence for the drift in demand rates and recommended estimation techniques for item variance-to-mean ratios as a function of time and mean demand.

Finally, we consider items whose failure is dominated by wear out phenomena. Here the variance is less than the mean, but the binomial distribution can be used to model the number of demands in a period.

4.2 Poisson Process

The exponential distribution of time until the next demand is given by:

$$\exp(t) = me^{-mt} \qquad t \geq 0 \tag{4.1}$$

which has mean $1/m$ and a variance of $1/m^2$. Note that m was defined as the average annual demand before, so $1/m$ is the average time between demands, measured in years. We want to show that if the time between demands has an exponential distribution, then the probability distribution for the number of demands in any specified time period, T, is Poisson.

The probability of no demands in time T is the probability that the next demand occurs after T:

$$\int_T^\infty me^{-mt}\,dt = e^{-mT} = p(0 \mid mT) \qquad T \geq 0 \tag{4.2}$$

where the last equality is Equation 2.4. Now we show that if the probability of x demands in any time t is Poisson for any x and the time between demands is exponentially distributed, the probability of $x + 1$ demands in any time T, where $T \geq t$, is Poisson.

$$\int_0^T p(x \mid mt)mte^{-m(\tau-t)}e^{-m(T-\tau)}dt = (m^{x+1}e^{-mT} / x!)\int_0^T t^x dt$$

$$= (mT)^{x+1}e^{-mT} / (x+1)! = p(x+1 \mid mT) \qquad T \geq 0 \tag{4.3}$$

The first expression in Equation 4.3 is the Poisson probability of x demands in time t, one demand at time $\tau \geq t$, and no demands during the final $(T - \tau)$. Since we have shown that the exponential time between demands leads to Poisson state probabilities for $x = 0$, and that if it is true for any arbitrary x it is true for $x + 1$, the result is true for all x by induction.

Now we want to consider the probability distribution for the time to the next demand, t, conditional on the fact that there has been no demand up to some time T. This is just the probability that the time to the next demand is t, divided by the probability that there has been no demand through time T:

$$me^{-mt} / e^{-mT} = \exp(t) / [1 - \int_0^T \exp(t) dt] = me^{-m(t-T)} \qquad t \geq \mathrm{T} \qquad (4.4)$$

This is still an exponential distribution with a new time origin of T instead of zero. This is the reason that the exponential distribution is called "memoryless", and why it is so useful in modeling. The demand rate at time T is obtained by substituting that value for t in Equation 4.4, showing that the demand rate is a constant, m, for any arbitrary T.

It is important to distinguish between the demand rate, the probability distribution of time to the next demand, and the probability distribution for the number of demands during an arbitrary period of time.

4.3 Negative Binomial Distribution

The negative binomial distribution is given by:

$$\mathrm{neg}(x) = \binom{a+x-1}{x} b^x (1-b)^a \qquad x = 0, 1, 2 \ldots \qquad (4.5)$$

where $a > 0$ and $0 < b < 1$. In elementary statistics courses the negative binomial is frequently introduced as the probability that it takes $a + x$ trials to achieve exactly a successes where each trial has a probability of success equal to $(1 - b)$. This is just the binomial probability for $a - 1$ successes in the first $a + x - 1$ trials times the probability that the next trial is a success.

The first term on the right-hand side of Equation 4.5 is called the number of combinations of $a + x - 1$ things taken x at a time. When a is integral the number of combinations equals $(a + x - 1)!/[x!(a - 1)!]$. The justification is that the number of sequences of successes and failures is $(a + x - 1)!$. But each sequence contains x successes in specific locations, and these can be rearranged in $x!$ ways that are indistinguishable. Similarly the $a - 1$ failures can be rearranged in $(a - 1)!$ ways that are indistinguishable. The reader who is not familiar with combinations is referred to Problem 1 at the end of this chapter.

By dividing factorials the number of combinations can be written

$$\binom{a+x-1}{x} = \frac{(a+x-1)(a+x-2)...(a)}{x!} \qquad x = 0, 1, 2 \ldots \qquad (4.6)$$

which is well defined and easily computed even when a is non-integral. From Equation 4.5 it is clear that x must be a non-negative integer.

Using the definitions in Equations 2.1 and 2.2, the mean, μ, and variance-to-mean ratio, V, are found to be

$$\mu = ab/(1 - b) \qquad V = 1/(1 - b) \tag{4.7}$$

(See Problem 3). From the formula for V it is clear that it must exceed one. Since there are two parameters to the negative binomial, we can derive the parameters a and b to represent any m and any V greater than one. After a little algebra, we find:

$$a = \mu/(V - 1) \qquad b = (V - 1)/V \tag{4.8}$$

This will be the most useful form for the parameters, because we want to be able to take an observed mean and variance-to-mean ratio (greater than one), calculate the parameters a and b, and generate the probability distribution. As in the Poisson, it will turn out that the most convenient computational procedure is to calculate neg(x + 1) by recursion from neg(x) (see Problem 4).

As b approaches 0, V approaches one as for the Poisson. In fact, it can be shown that the negative binomial approaches the Poisson probabilities as b gets smaller, even though the negative binomial can never have a $V = 1$ (see Problem 5). This is important, because we would not want a discontinuity between any Poisson probability, $p(x)$, and the corresponding negative binomial probability, neg(x), as b approaches zero.

In summary the important characteristics of the negative binomial are:

- Discrete distribution for non-negative arguments
- Generalization of the Poisson
- Distribution with two parameters that are easy to estimate for a specified mean and variance-to-mean ratio greater than one
- Recursion formulas that are easy to compute

In the previous section we showed that Poisson state probabilities for the number of demands in any time period T are equivalent to an exponential distribution for the time between demands. This is what we mean by a *Poisson process.* It is natural to suppose that we will now consider the distribution of time between demands that is equivalent to a negative binomial distribution for state probabilities.

Such a distribution for the time between demands does exist. It consists of an exponential distribution for the time between clusters of demand, and a logarithmic distribution for the number of demands in each cluster. It is called a *logarithmic Poisson process,* a member of the compound Poisson family of processes, which are generalizations of the Poisson process. Much of our research in the 1960s assumed compound Poisson demand, because we kept observing variance-to-mean ratios greater than one in our data.

If demand were compound Poisson, the variance-to-mean ratio for an item would be constant, regardless of the length of the time period over which demand is observed. Instead, the empirical data that we discuss below show variance-to-mean ratios that increase with time. Also, there would be clusters of demand, and this is not observed either. For these reasons we will not discuss the compound Poisson further in this book. The reader who is interested in pursuing the compound Poisson should see Feller (1958).

However, there is an important fact about the compound Poisson shown by Feller that we will use. The compound Poisson is the most general "independent increments" process, by which we mean that the number of demands in two non-overlapping time periods is independent. In other words for any non-negative values of y (discrete), μ_1, and μ_2:

$$\Pr\{\mu_1 + \mu_2\} = \sum_{x=0}^{y} \Pr\{x|\mu_1\}\Pr\{y - x|\mu_2\} \tag{4.9}$$

where x demands are observed in time period 1 with mean μ_1, and y - x demands are observed in period 2 with mean μ_2. The practical implication of this is that if demand over any time period t_1 has a negative binomial distribution with variance-to-mean ratio V, and demand has a negative binomial distribution over another time period t_2 with the same variance-to-mean ratio, then the probability distribution over $t_1 + t_2$ is negative binomial with that variance-to-mean ratio and a mean that is the sum of the two period means.

This is true for the negative binomial, the Poisson, and the binomial (which is described below in Section 4.17). It is most easily shown with generating functions, which we have not used in this book. But the property can be demonstrated arithmetically as in Problems 12-14 at the end of the chapter.

The negative binomial will be used below to represent the probability distribution of demand over a specified time period. As the time period lengthens, the variance-to-mean ratio will tend to increase. This is not consistent with a compound Poisson process, as noted above, but it is consistent with a Poisson process with non-stationary increments.

4.4 Multi-Indenture Problem

As noted in the overview, one of the reasons we use the negative binomial is to model the mean and variance for the pipelines where the variance exceeds the mean. We will illustrate this by showing how the accuracy of the METRIC theory can be improved in the multi-indenture case.

The multi-indenture problem for a single base was addressed by Sherbrooke (1971), and for multi-echelon by Muckstadt (1973) in a model called MOD-METRIC. In this section we discuss the single base case with two indentures.

When a demand for a first-indenture, line-replaceable unit (LRU) occurs at the base, a spare LRU is issued if one is on hand. If no spare LRU is available at the base, a backorder is established. In either case the LRU is repaired at the base, and during this process one second-indenture, shop-replaceable unit (SRU) is identified as having failed. If a spare SRU is available, the failed SRU is removed and replaced and the LRU repair is completed; otherwise an SRU backorder is established. In either case the SRU is repaired at the base.

The important assumptions are that the SRU repairs are not delayed for spare parts, and that each LRU failure is due to one and only one SRU failure. The first assumption applies only to the lowest indenture level considered, and thus becomes less important as more indentures are modeled. The second assumption has been discussed at some length by Muckstadt (1982). Approximations for the case of multiple SRU failures have been developed by Sherbrooke (1988).

We will use the subscript i to indicate the item as before, with $i = 0$ for the LRU and $i > 0$ for the SRUs. The subscript $i = 0$ should not be confused with $j = 0$ in the previous chapter, which indicated the depot. (In Chapter 5 we will need both subscripts for VARI-METRIC.)

If the demand process for the base is Poisson with average annual demand m_0, the demand rate for SRU i must be Poisson with average annual demand m_i and

$$m_0 = \sum_{i=1}^{I} m_i \tag{4.10}$$

Assume that the repair times T_0 for the LRU and T_i for the SRUs are constant. Then $x_0(t)$, the number of LRUs in repair at a random point in time t can be written

$x_0(t)$ = LRU demands during the time interval $(t - T_0, t)$

$+ \sum_{i=1}^{I}$ LRU demands prior to t-T_0 that required SRU i to repair, but SRU i is still backordered at time t (4.11)

The number of LRUs in repair equals the demand in the time period $(t - T_0, t)$ and this is independent of the LRU delays due to SRU backorders because they arose from LRU demands in an earlier segment of the Poisson process prior to $t - T_0$. Furthermore, the LRU delays due to different SRUs are independent of each other, because of our assumption that each LRU failure is due to one and only one SRU failure and because a probabilistic split of a Poisson process results in independent Poisson processes.

Thus, the expected pipeline for the LRU can be written:

$$\mathrm{E}[X_0] = m_0 T_0 + \sum_{i=1}^{I} \mathrm{EBO}(s_i | m_i T_i) \tag{4.12}$$

and the expected LRU backorders at base are $\mathrm{EBO}(s_0 \mid x_0)$.

It is easy to generalize this result from constant repair times to arbitrary probability distributions of repair. Consider the SRUs first. Since the input to each SRU repair process is Poisson and independent, Palm's theorem can be applied.

Now consider the LRU and suppose that there are two possible repair times $T_0(1)$ and $T_0(2)$ and corresponding probabilities such that the overall average repair time is still T_0. The original Poisson process is divided into two independent Poisson processes, and the arguments of the previous paragraph still apply. This result can be extended by induction to any discrete repair distribution.

For continuous repair distributions we can use the queueing analogy where the LRU and any SRU correspond to two infinite channel queues in series. The number of units of the LRU in repair is Poisson from Palm's theorem. The output of the first queue is Poisson, since the input was Poisson and the repair process is independent of the input.[1] Since the output of the first queue is the input to the second, we can apply Palm's theorem again.

[1] This is a special case of a result of Erlang proved in Syski (1986), but a formal mathematical proof is not required. The inductive argument above is sufficient here.

4.5 Multi-Indenture Example

In this section we want to take the multi-indenture result for the expected LRU backorders at base and compare its accuracy with a simulation of the multi-indenture problem.

Consider an example such that when the LRU fails, there is a .5 probability that it is because of SRU 1 and a .5 probability of SRU 2. We assume an LRU demand rate of 2/day, stock levels and average repair times as shown in Table 4-1, from which we calculate the average number of units of the LRU in repair. Substituting these values into Equation 4.12 yields the expected pipeline for the LRU:

$$E[X_0] = 1 + EBO(10|8)] + EBO(10|8)] = 1.852 = \mu_0$$

and the expected LRU backorders at base are:

$$EBO(s_0|\mu_0) = EBO(4|1.852) = .056.$$

Table 4-1. Multi-Indenture Example

Item	Stock Level	Demand Rate	Average Repair Time (T)	Average # in Repair
LRU	4	2/day	0.5 days	1
SRU #1	10	1/day	8 days	8
SRU #2	10	1/day	8 days	8

However, a simulation of this problem with 50,000 simulated years of experience shows that the expected LRU backorders are .202, with a 95% confidence interval of ±.005. Thus, in this example the MOD-METRIC procedure understates the expected backorders by nearly a factor of four.

4.6 Variance of the Number of Units in the Pipeline

The difference between the MOD-METRIC solution and the simulation is very large. In fact, the difference is atypically large in the example, because we purposely chose a combination of demand rates and stock levels that would produce a dramatic difference.

Why does this difference arise? When we compute the expected LRU backorders, based on the pipeline x_0, we are assuming that the number of units in the pipeline has a Poisson distribution. In fact, the number of units in the pipeline has a variance-to-mean ratio greater than one, as we shall now show.

In Equations 2.1 and 2.2 we provided formulas for the expected value (mean) and variance for any demand distribution. Then in Equation 2.7 the expected value of backorders, EBO(s), was defined, where this expectation was a function of the demand distribution and the stock level, s.

Now we need to estimate the variance of the backorder function, VBO(s), where we use the abbreviated notation discussed in the footnote following Equation 2.7. The formula is obtained by substituting the definition of backorders in that footnote into Equation 2.2:

$$\mathrm{VBO}(s) = \mathrm{E}[\mathrm{B}^2(X|s)] - [\mathrm{EBO}(s)]^2 \tag{4.13}$$

We already know how to compute the last term on the right of this equation as it is the square of Equation (2.7). The first term on the right is the second moment of the backorder function which is analogous to Equation (2.3) as applied to Equation (2.7):

$$\mathrm{E}[\mathrm{B}^2(X|s)] = \sum_{x=s+1}^{\infty}(x-s)^2 \Pr\{X = x\} \tag{4.14}$$

Note that when $s = 0$, Equation 4.14 reduces to Equation 2.3 and Equation 4.13 reduces to Equation 2.2. When the probability distribution is Poisson, the ratio VBO(s)/EBO(s) equals one for $s = 0$, but for any positive value of s the ratio exceeds one as shown by Svoronos (1986). The typical behavior is for the ratio to increase as a function of s to a maximum at a value of s slightly larger than the mean and then decrease asymptotically to one, as shown in Figure 4-1.

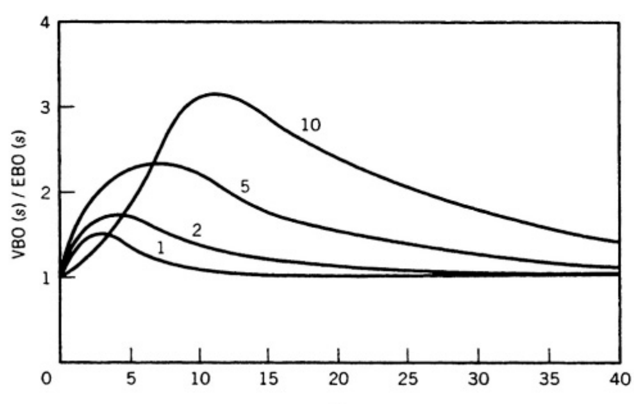


Figure 4-1. **VBO(s)/EBO(s) for various mean values of the Poisson**

Thus, even though the LRU demand is Poisson and the SRU demand is Poisson, the probability distribution of backorders for SRUs is not Poisson unless each SRU stock level $s = 0$. The pipeline for the LRU is the composition of the SRU backorder distributions, so it is Poisson only when each SRU stock level s = 0, also.

A comparison of the Poisson and negative binomial distributions is shown below in Table 4-2 for a mean of one. The Poisson has a variance-to-mean ratio of one always; the negative binomial we have chosen has a variance-to-mean ratio of 3. Thus, the negative binomial parameters from Equation (4.8) are

$$a = \mu/(V - 1) = 1/(3 - 1) = .5$$

and

$$b = (V - 1)/V = 2/3.$$

Since the means of the two distributions are each one but the variance-to-mean ratio of the negative binomial is three, the latter probabilities must be larger both near 0 and for large values of *s*.

We have shown the mean, variance, and variance-to-mean ratio of the backorders for both distributions as well. Note that the mean backorders, not just the variance, are larger for the negative binomial than for the Poisson at each stock level, *s*, greater than zero. Also, the variance-to-mean ratio of backorders for the negative binomial remains high for values of *s* well in excess of the mean.

For computational purposes it is useful to have recursion formulas. The reader was asked to derive the recursion formula for EBO(s) in Problem 3 of Chapter 2. The following formula for the second moment was originally derived by Michael Konvalinka of the Logistics Management Institute and is easily verified (see Problem 6):

$$\mathrm{E}[\mathrm{B}^2(X|s)] = \mathrm{E}[\mathrm{B}^2(X|s-1)] - \mathrm{EBO}(s) - \mathrm{EBO}(s-1) \qquad s>0 \qquad (4.15)$$

Note that these recursion formulas for the mean and second moment of backorders are valid for any state probability distribution, not just the Poisson. Now we are in a position to compute the variance of the pipeline, which looks very similar to the formula for the expected value from Equation 4.12:

$$\mathrm{Var}[X_0] = m_0 T_0 + \sum_{i=1}^{I} \mathrm{VBO}(s_i | m_i T_i) \qquad (4.16)$$

The justification for this formula is straightforward. Each term on the right-hand side is independent of each other, and the variance of a sum of independent variables is the sum of the variances. From the discussion at the

Table 4-2. **Poisson and Negative Binomial Distributions with Mean = 1**

Stock	Poisson*				Negative Binomial†			
(s)	p(s)	EBO(s)	VBO(s)	VBO(s)/EBO(s)	neg(s)	EBO(s)	VBO(s)	VBO(s)/EBO(s)
0	.3679	1.0000	1.0000	1.0000	.5774	1.0000	3.0000	3.0000
1	.3679	.3679	.4968	1.3504	.1925	.5774	2.0893	3.6188
2	.1839	.1036	.1499	1.4460	.0962	.3472	1.3776	3.9684
3	.0613	.0233	.0331	1.4176	.0535	.2132	.8924	4.1861
4	.0153	.0043	.0059	1.3619	.0312	.1327	.5744	4.3298
5	.0031	.0007	.0009	1.3109	.0187	.0833	.3691	4.4290
6	.0005	.0001	.0001	1.2633	.0114	.0527	.2372	4.5002
7	.0001	.0000	.0000	1.1455	.0071	.0335	.1526	4.5533
8	.0000				.0044	.0214	.0984	4.5942
9	.0000				.0028	.0137	.0635	4.6268
10	.0000				.0018	.0088	.0411	4.6534
11	.0000				.0011	.0057	.0266	4.6758
12	.0000				.0007	.0037	.0173	4.6952
13	.0000				.0005	.0024	.0112	4.7123
14	.0000				.0003	.0015	.0073	4.7280
15	.0000				.0002	.0010	.0047	4.7429
16	.0000				.0001	.0007	.0031	4.7580
17	.0000				.0001	.0004	.0020	4.7745
18	.0000				.0001	.0003	.0013	4.7942

* Variance = 1 of Poisson.
† Variance = 3 of negative binomial.

end of Section 4.4, we know that even when there is a probability distribution for repair time, the number of units of the LRU in repair is Poisson and the variance of a Poisson equals the mean.

4.7 Multi-Indenture Example Revisited

Let's calculate the pipeline variance for our example, whose data were given in Table 4-1:

$$\mathrm{Var}[X_0] = 1 + \mathrm{VBO}(10|8) + \mathrm{VBO}(10|8) = 3.468.$$

The pipeline variance-to-mean ratio, V = 3.468/1.852 = 1.873, so the Poisson distribution with a variance-to-mean ratio of one is clearly a poor approximation for the number of units in the pipeline. Suppose we substitute the pipeline mean and variance-to-mean ratio into Equation (4.8) to estimate the parameters of a negative binomial. Then we find that

$$\mathrm{EBO}(s_0|\mathrm{E}[X_0], \mathrm{Var}[X_0]) = \mathrm{EBO}(4|1.852, 3.468)] = .194$$

where the notation has been expanded to show the conditioning on the mean and variance of the state probability distribution. This estimate of the expected base backorders is a lot closer to the simulated value of .202, even though it does fall outside the 95% confidence limits of .197 - .207. However, we selected stock levels for the example to produce an atypically large difference between MOD-METRIC and simulation. In most cases the agreement between them will be closer.

The negative binomial was employed by Michael Slay and later by Steve Graves and this author. How do we know that the negative binomial is the right probability distribution to use? The answer is that we don't know the right distribution, but we do know that a probability distribution whose mean and variance agree with the values for the pipeline will do a lot better than the Poisson. Even though the probability distribution we use is somewhat different from the true distribution, our interest is not the probabilities themselves, but in a function of these probabilities. And the difference between the values of the true and approximate function will be smaller then the corresponding differences between the individual probabilities.

We should include a historical note that the VARI-METRIC improvements were first made in the multi-echelon problem, rather than in the multi-indenture problem. For pedagogical reasons the VARI-METRIC idea is introduced here in the simpler multi-indenture context. VARI-METRIC for the combined multi-echelon, multi-indenture problem is developed in Chapter 5.

In summary, the negative binomial has several properties that recommend it for our purposes: 1) discrete distribution for non-negative arguments; 2) generalization of the Poisson; 3) distribution with two parameters that are easy to estimate for a specified mean and variance-to-mean ratio greater than one; 4) recursion formulas that are easy to compute.

4.8 Demand Rates that Vary with Time

The second reason for employing the negative binomial is to model Poisson demand processes with non-stationary increments (where the mean may drift over time). Slay and Sherbrooke (1988) showed that over short periods of time demand follows a Poisson process with a constant mean for most items. Sherbrooke (1984) showed that the variance-to-mean ratio for an item tends to increase as the time period over which demand is measured becomes longer. The only model consistent with these observations is a Poisson process with non-stationary increments. Let's consider a Poisson process with annual mean m. But, during half of the year the mean is $m + M$ and during the other half it is $m - M$ (where $m > M$), so that the demand rate is not constant during the year. Our objective is to see what happens to the variance-to-mean ratio over the year.

The variance over the entire year is computed from Equation 2.2, using the fact that the second moment is the square of the mean plus the variance for each of the two parts of the Poisson process:

$$\begin{aligned} \mathrm{Var}[X] &= \mathrm{E}[X^2] - [\mathrm{E}(X)]^2 \\ &= .5[(m+M)^2 + m + M] + .5[(m-M)^2 + m - M] - m^2 \\ &= .5[m^2 + 2mM + M^2 + m + M + m^2 - 2mM + M^2 + m - M] - m^2 \\ &= M^2 + m \qquad (4.17) \end{aligned}$$

The first term is the second moment of the Poisson process with mean $m + M$ for .5 year, the second term is the second moment of the Poisson process with mean $m - M$ for .5 year, and the third is the square of the mean for the entire year. Since the mean is m, the variance-to-mean ratio over the entire year is greater than one for any $M > 0$.

Note that the result in Equation 4.17 is unaffected if the demand rate changes many times between $m + M$ and $m - M$, provided only that when totaled the amount of time with each mean is the same. Dividing Equation 4.17 by the mean, m, we see that the variance-to-mean ratio is $M^2/m + 1$. Since $m > M$, the maximum variance-to-mean ratio is $m + 1$, which implies that the Poisson assumption is likely to be a more significant error for large values of m.

In Table 4-3 below we show the behavior of the variance-to-mean ratio for a few values of m and M. For a specific value of M, the impact on variance-to-mean ratio decreases with m. However, when M is a fixed percentage of m, which is probably more meaningful, the impact on variance-to-mean ratio increases with m. (As an example of the latter, where $M = .5m$, compare the case with $m = 1$, $M = .5$, leading to a variance-to-mean ratio of 1.25 and the case with $m = 10$, $M = 5$, leading to a variance-to-mean ratio of 3.50).

Table 4-3. Variance/Mean Ratio as a Function of m, M in Example (Section 4.8)

m	M	Variance/Mean
1	.5	1.25
1	1	2
10	.5	1.025
10	1	1.10
10	5	3.50
10	10	11.00

We do not claim that mean demands change in "quantum leaps" from one level $M + m$ to $M - m$ as above, but it is clear that any change in mean demand for a Poisson process results in variance-to-mean ratios greater than one for the number of demands in a given time period (the state probabilities). Furthermore, these variance-to-mean ratios are likely to be greater when the mean is larger, a fact of some importance that we will encounter again below.

4.9 Bayesian Analysis

Our objective is to estimate how these means of different items change or drift over time. For this purpose we will describe some demand prediction experiments, but first we want to introduce Bayesian analysis as one of the candidate prediction techniques.

In the section below on Bayesian analysis and in several subsequent sections we will use conditional probabilities. Because of their importance it is appropriate to devote a brief section to them now.

Consider two events a and b that are not independent. From elementary probability theory we can write

$$\Pr\{ab\} = \Pr\{a|b\}\Pr\{b\} = \Pr\{b|a\}\Pr\{a\} \tag{4.18}$$

which states that the joint probability of both a and b equals the conditional probability of a given that b occurs multiplied by the probability that b

occurs. The joint probability also equals the conditional probability of b given that a occurs multiplied by the probability that a occurs.

Consider the following inference problem. Assume that as a part of a routine screening process, your physician tells you that you have tested positive for cancer. He tells you that the test isn't perfectly accurate - there is a 2% false positive rate (a healthy person will be diagnosed as having cancer 2% of the time) and a 2% false negative rate (a person with cancer will fail to be detected 2% of the time). How depressed should you get? And what data do you need to evaluate the problem dispassionately?

It turns out that this is an example of a problem that can be systematically analyzed by using Bayesian analysis, named for the Reverend Thomas Bayes, a British clergyman whose theorem was published posthumously in 1763. Bayes' rule is obtained from Equation 4.18 after division on both sides of the right-hand equality by the probability of b

$$\Pr\{a \mid b\} = \frac{\Pr\{b \mid a\}\Pr\{a\}}{\Pr\{b \mid a\}\Pr\{a\} + \Pr\{b \mid \bar{a}\}\Pr\{\bar{a}\}} \tag{4.19}$$

where we have expanded the denominator for the probability of b into two terms: (1) where b occurs given a times the probability of a; (2) b occurs given a does not occur (as indicated by the line over the a) times the probability that a does not occur. More generally when a can assume more than two values, the denominator is a sum over all possible states of a.

To analyze the medical inference problem we need to define the events a and b. Since the objective is to assess the probability of cancer given a positive test result we assign

a = cancer $\qquad\qquad \bar{a}$ = no cancer

b = positive test result

The problem data has already supplied some of the values we need:

$$\Pr\{b|a\} = \Pr\{\text{positive test}|\text{cancer}\} = .98$$

$$\Pr\{b|\bar{a}\} = \Pr\{\text{positive test}|\text{no cancer}\} = .02$$

But, it is clear that we need the probability that an individual tested in a random screening has cancer, Pr{a}. Suppose that it is known from previous screenings that this averages about .005. Then substituting into Equation 4.19 yields:

$$\Pr\{\text{cancer}|\text{positive test}\} = (.98)(.005)/[(.98)(.005)+(.02)(.995)]$$

$$= .198$$

It seems surprising to individuals unfamiliar with Bayes' rule that with only a 2% rate for false positives and false negatives, there is still only one chance in five that an individual who has tested positive has cancer. Of course, the results are critically dependent on the rate of cancer in the population being screened. This probability, $\Pr\{a\}$ is known as the prior probability, and the probability obtained after applying Bayes rule is known as the posterior probability (because these probabilities are before and after using the information from the screening test). In this example, the prior probability of having cancer of .005 is increased 40 times to the posterior value of .198, because of the positive test result. If the prior probability of cancer in the screening population were twice as large, .01, the probability of really having cancer, given a positive test result, jumps to .33.

In summary, the utility of Bayes' rule depends largely on whether there is reasonable agreement over what the prior probabilities are. If different analysts estimate the prior probabilities very differently, then the posterior probabilities will usually be very different as well. On the other hand, there are a number of problems where it is reasonable to assume that all prior states are equally likely (e.g. games of chance, submarine location).

How does all this relate to the estimation of demand rates and inventory modeling in general? In the next section we discuss an "objective" Bayes approach, where the idea is to estimate the prior distribution from data in a manner such that different analysts will get the same - or similar - results. In the succeeding section we address the non-Bayesians who don't believe in prior distributions of any kind, and show that classical statistics leads to James-Stein estimators, a Bayes-like procedure anyway.

4.10 Objective Bayes

Suppose that we have observed demand on all first-indenture items for a particular aircraft type at a base over a period of time, such as six months or a year. Let's assume that we don't have any initial demand estimates made prior to this period of flying, or any such initial estimates have little credibility now.

We demonstrate below a procedure for using the demand data on the group of items to estimate the parameters of a prior distribution. Then we use Bayes' theorem to combine the prior distribution on all items with a Poisson demand process to estimate a posterior distribution of demand for each individual item. The mathematically challenged reader may wish to skip the following derivation of the prior distribution parameters and rejoin us after Equation 4.26.

First, we rewrite Bayes' theorem from Equation 4.19:

$$h(\mu \mid x) = p(x \mid \mu)\text{gam}(\mu) / \int_0^\infty p(x \mid \mu)\text{gam}(\mu)d\mu \tag{4.20}$$

where the gam denotes a gamma distribution and p a Poisson with mean μ, and h is the posterior distribution for the mean of an item, conditional on the observed number of demands x. It should be clear that Equation 4.20 is the continuous variable analogue for Equation 4.19.

The gamma distribution is given by:

$$\text{gam}(\mu) = b^{-a}\mu^{a-1}e^{-\mu/b} / \Gamma(a) \qquad \mu > 0 \tag{4.21}$$

where the parameters $a > 0$, $b > 0$. The mean = ab, the variance-to-mean ratio = b. Note that $\Gamma(a)$ indicates the gamma function, not to be confused with the gamma distribution. For any value of $a > 1$ the gamma function has the property:

$$\Gamma(a) = (a - 1)\,\Gamma(a - 1) \tag{4.22}$$

When a is integral, the gamma function of a equals $(a - 1)$ factorial. For non-integral values of a, it is necessary to use a computer or mathematical tables.

The essence of objective Bayes is to estimate the parameters of the gamma prior distribution from the demand data. This is done by "moment matching," setting the mean and variance of the random variable X in the denominator of Equation 4.20 to the mean and variance of the observed period demand across the group of aircraft items. That will provide two equations for the two parameters of the gamma prior distribution, a and b. The expected value of the random variable X is the Poisson probability that $X = x$ for each possible x, multiplied by x, and integrated over the probability that the Poisson mean is μ:

$$E[X] = \sum_{x=0}^{\infty} x \int_0^\infty p(x|\mu)\text{gam}(\mu)d\mu = \int_0^\infty E[X|\mu]\text{gam}(\mu)d\mu$$

$$= \int_0^\infty \mu\text{gam}(\mu)d\mu = ab \tag{4.23}$$

This follows easily since the mean of the Poisson is μ, and the mean of the gamma is *ab.* Similarly,

$$E[X^2] = \sum_{x=0}^{\infty} x^2 \int_0^{\infty} p(x|\mu)\text{gam}(\mu)d\mu = \int_0^{\infty} E[X^2|\mu]\text{gam}(\mu)d\mu$$

$$= \int_0^{\infty} [\mu + \mu^2]\text{gam}(\mu)d\mu = ab + ab^2 + (ab)^2 \qquad (4.24)$$

where the second moment is the variance plus the mean squared. We compute the variance by substituting Equations 4.23 and 4.24 into Equation 2.2:

$$\text{Var}[X] = E[X^2] - [E(X)]^2 = ab(1 + b)$$

Note that $E[X]$ and $\text{Var}[X]$ are computed across the group of all items from their observed demand during the period. Thus the parameter b can be estimated from these data as

$$\text{Var}[X]/E[X] = ab(1 + b)/(ab) = 1 + b$$

$$b = \text{Var}[X]/E[X] - 1 \qquad (4.25)$$

The parameter a is estimated from Equation (4.23):

$$a = E[X]/b \qquad (4.26)$$

Let's summarize what we have accomplished. We have found a way to estimate the two parameters, a and b, of the best-fitting gamma prior distribution from the mean and variance of observed demand across all items over a six-month or one-year time period. The prior distribution is the same for all items.

Then a posterior distribution is computed for each item based on its observed demand, x, during the period. Bayes' theorem from Equation 4.20 is employed. Note that all items with the same observed demand, x, have the same posterior distribution.[1]

[1] The negative binomial distribution is ubiquitous. The denominator of Equation 4.20, the probability distribution for a Poisson variable with a gamma prior distribution, has a negative binomial distribution. This is not terribly important, but can be shown easily (Problem 15).

It turns out that this posterior distribution is gamma as well where the prior parameter a becomes $a + x$ for the posterior and the prior parameter b becomes $b/(b + 1)$. (see Problem 9).

Figure 4-2 gives a graphical illustration of how Bayes' theorem operates on the prior distribution. If the observed value of x exceeds the prior mean (ab), the posterior distribution is shifted up toward larger demand rates with a mean $(a + x)b/(b + 1)$ that exceeds ab. But, the posterior mean is less than the observed value x.

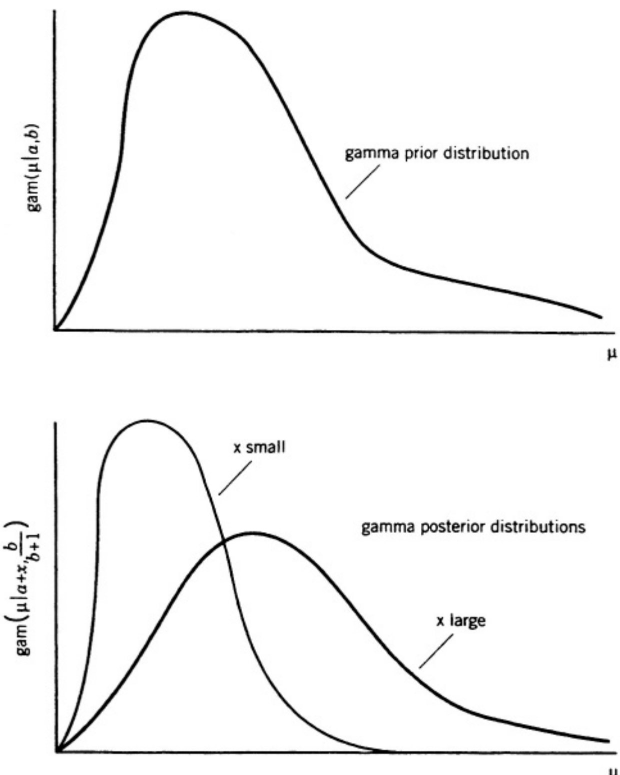


Figure 4-2. Bayes' procedure

Similarly an observed value of x less than the mean results in a posterior distribution whose mean lies between the observation, x, and the prior mean, ab. Regardless of the value of x we can say that the observed value x is "regressed toward the (prior) mean" in order to obtain the posterior mean.

What is the real value, if any, of this procedure? Usually an item with no demand during the period would be assumed to have a zero demand rate. But, we don't really believe that zero is the best estimate of the demand rate for every item that has no demand during six months or a year. The true demand rates are probably low, but we know from experience that if we observe this group of zero-demand items for another time period, some will have positive demand. The Bayes procedure will result in positive, though small, demand rates for these zero demand items.

The use of zero-demand-rate estimates for all items in the group becomes even less sensible if the flying hours or the number of aircraft is increasing over the next period. Suppose that the number of flying hours is to increase by a multiple of k greater than one, and flying hours are thought to be a primary determinant of demand. More of the items with zero observed demand in the past period are likely to have positive demand during the period of greater flying. The traditional approach has been largely judgmental (the logistician estimates positive demand rates for some items), but in the Bayesian approach we know that an item with gamma parameters a and b would be converted to an item with gamma parameters a and kb. This provides a natural mechanism to reflect changes in the flying program (up or down).

Similar considerations apply to items with very large observed demand. The Bayes procedure tends to decrease the mean demand rate estimates for the items with very large observed demand.

For those who wonder if this technique is merely of academic interest, we point out that this is actually the technique used in the George Air Force Base Field Test of Chapter 1 that was described in Air Force Logistics Command (1967). The "objective Bayes" technique was reported on in Feeney and Sherbrooke (1965).[1] The consideration of unit cost in conjunction with a probability distribution for demand led to positive stock levels under the new policy for some items that had no demand during the previous six months. Items with no demand and a high cost would not be stocked by either policy.

It should be noted that the "objective Bayes" procedure bears some similarity to empirical Bayes which has been studied at length by Robbins (1964) and many others. An extensive bibliography is provided in Sherbrooke (1969). The empirical Bayes problem arises when we have a sequence of pairs of random variables $(\mu_1, x_1), (\mu_2, x_2), \ldots, (\mu_n, x_n)$, each pair being independent of all the other pairs, the μ_i having a common prior distribution, and the conditional distribution of x_i given that $\mu_i = \mu$ being specified by a known probability distribution $\Pr\{X_i = x_i | \mu\}$. The values μ_i are not observed and the prior distribution is unknown. Based on an additional observation x_{n+1}, we want to make an optimal decision about μ_{n+1}. It turns out that in some problems, the empirical Bayes procedure with no information about the prior converges as $n \to \infty$ to the optimal Bayes decision

[3] The log normal distribution was used as the prior in that report, but the results are very similar to those that would be obtained with the gamma prior.

if the prior distribution were known and the the optimal Bayes decision based on the observation x_{n+1} were made.

The difference is that in objective Bayes we try to estimate the prior distribution; in empirical Bayes we never estimate the prior itself, but the (stockage) decisions that would be made if we knew the prior. The drawback with empirical Bayes is the large amount of data that is required for convergence of the decision rules to optimality. In our application empirical Bayes would be particularly suspect, because the underlying demand rates seem to change with time.

4.11 Bayesian Analysis in the Case of Initial Estimate Data

Before leaving the topic of Bayes completely, note that when there are both initial estimates and demand data by item it is possible to define a different prior for each item and use Bayes theorem. A number of authors have used this procedure. The initial estimate anchors the prior mean. But, if the prior distribution has two parameters like the gamma, some subjective estimate of the prior variance on each item is still required.

Let's illustrate this procedure with an example. Suppose that we have observed demand over some number of operating hours. When we multiply the number of operating hours, t, by the initial estimate per hour, assume that the expected demand is $\mu = 20$. However, the observed demand, $x = 10$.

Now if we believe that the initial estimate and the observed data over t hours contain about the same amount of "information", it would be natural to average the two estimates and obtain a best estimate of demand = 15.

As noted in the previous section, the mean of the gamma prior is ab so one equation for the best-fit gamma distribution parameters would be ab = 20. We also noted that when the observed demand, x, has a Poisson distribution, the posterior distribution is gamma as well where the prior parameter a becomes $a + x$ for the posterior and the prior parameter b becomes $b/(b + 1)$ - see Problem 9.

Thus when $x = 10$, the posterior mean must satisfy $(a + 10)b/(b + 1) = 15$. for the posterior mean. To solve for a and b we substitute the first equation, $a = 20/b$, into the second, obtaining:

$$(20/b + 10)b = 15b + 15$$

$$20 + 10b = 15b + 15$$

$$5 = 5b$$

or $b = 1$, $a = 20$. As noted just above Equation 4.25, the variance-to-mean ratio of the prior, $V = 1 + b$. Since b becomes $b/(b+1)$ in the posterior, it is clear that the variance-to-mean ratio must decrease for any value of b; or, to put it another way, as we use Bayes theorem over and over we become ever more certain of our estimate. But, this is at odds with our belief that demand rates change. Thus, it is important to restrict the time period over which we apply Bayes.

Of course, if the observed data is thought to contain twice as much "information", say, as the initial estimate, then the second equation for the posterior mean would be $(a + 10)b/(b + 1) = 13.3$. We recommend that the analyst attempt to establish the value of t for equal "information" before collecting data. There may be reasons to revise this later, but it is important not to let the observed value of x influence the selection of t. Of course, the same value of t may not be appropriate for all items; it should be smaller for items whose technology is "pushing-the-state-of-the-art" and for which initial estimates are "wild-ass guesses" (WAGs).

We reviewed a large amount of early program data from the Air Force as members of the RAND Corporation Logistics Department in the 1960s. Our conclusion was that the Air Force used initial estimate data for a longer time period than warranted at which point they would switch to observed data. In the large majority of cases a better estimate would have been to use the initial estimates and modify them with Bayes, giving progressively more weight to the observed data. Of course, at some point the initial estimates become meaningless, and as we have suggested earlier it is desirable to weight recent data more heavily if we believe demand rates change.

4.12 James-Stein Estimation

There are individuals who believe that there is too much subjectivity with prior distributions to use Bayesian analysis, despite empirical and objective Bayes procedures. They insist that classical statistics must be used for inference. For example, in the estimation of a baseball player's "true" batting average, classical statistics can be used to show that the observed average is uniformly better than any other estimator. (In the language of the statistician, which we shall not bother to explain in detail here, the average is the maximum likelihood, minimum variance, unbiased, consistent, efficient estimator). The paradoxical element in James-Stein theory is that if we have three or more baseball players and we are interested in predicting future batting averages for each of them, there is a better procedure than simply extrapolating from the three separate averages. This example and a particularly lucid, elementary description of James-Stein estimators are provided by Efron and Morris (1977).

It turns out that each observed batting average, x_i, for player i should be shrunk toward the overall mean $\bar{x}$ to obtain an estimate y_i:

$$y_i = \bar{x} + \kappa(x_i - \bar{x}) \tag{4.27}$$

where κ is a "shrinking" constant in the range [0,1] to be determined below. The expected value of the sum of the squared errors between the y_i and the "true" batting average for each player is never more than the corresponding sum for the x_i's, and is usually substantially less.

For their example Efron and Morris picked all major league baseball players in the 1970 season who had batted exactly 45 times on the day the data were tabulated. There were 18 such players with observed averages ranging from .400 to .156. The average for the 18 players, $\bar{x}$, was .265 and the shrinking constant κ =.212. Thus, application of the formula to Roberto Clemente with the highest observed average of .400 results in an estimate for the season from Equation 4.27 of .265 + .212(.400-.265) = .294. A player near the other extreme such as Thurman Munson with an observed average of .178 receives a James-Stein estimate of .247.

How well did the James-Stein estimator do when evaluated on the "true" batting averages which we construe as the season averages here? The James-Stein estimator was closer to the season average for 16 of the 18 players, and the sum of squared errors was .022 as contrasted with a squared error sum of .077 from the original observed averages. It can be shown that the improvement in squared error by a factor of about 3.5 is not just luck, but close to the expected ratio of the squared errors.

The shrinking constant is critical to the James-Stein procedure. One formula for κ is:

$$\kappa = 1 - \frac{(I-3)\sigma^2}{\sum_{i=1}^{I}(x_i - \bar{x})^2} \tag{4.28}$$

where I is the number of unknown means (players), σ^2 is the variance in a player's batting ability[1], and $\sum_{i=1}^{I}(x_i - \bar{x})^2$ is the sum of the squared deviations of the individual averages x_i from the grand average $\bar{x}$. (In a logistics application I might be the number of items.)

[1] This is a slight simplification, because a change of variable is made to stabilize the variances of different players.

We need an estimate of the variance in a player's batting average. The binomial distribution, described in Section 4.17 below, can be used to estimate the number of base hits by a player in n at-bats. The parameter p is the probability that a player gets a hit, $np(1-p)$ is the variance of the binomial and $p(1-p)/n$ is the variance in the observed mean (batting average). It is this latter quantity that we substitute for σ^2 in Equation 4.28, using estimates of. $p = .265$ and $n = 45$. For a fixed I and σ^2, as the dispersion of the x_i increases around $\bar{x}$, the denominator increases and κ becomes larger. This in turn implies that the observed averages will be shrunk less, because the individual averages are less similar.

As the season progresses (the overall season might comprise 400 or more at-bats), both the variance in the numerator of Equation 4.28 and the dispersion of the x_i around $\bar{x}$ will decrease. The value of κ will tend to increase so that less weight is given to the overall average as more data accumulates on each player.

4.13 James-Stein Estimation Experiment

As an illustration of James-Stein we wrote a simulation with ten items. The application we have in mind is the space station (see Chapters 6 and 7) where at any point in time the number of years of experience on each item will vary, because the number of units of each item on the station will differ and items are installed at different times in the station construction.

We provide the simulation with constant demand rates for a Poisson process and then use a random number generator to draw the number of "observed" demands as shown in Table 4-4. For example, the average annual demand for item 1 is 0.800, and the simulation generates 4 demands in 4 years. The "usual" estimate is the demand divided by the number of years, namely 1.000. While the mean is an unbiased, consistent, maximum likelihood estimator, we know already that it is not the most efficient when there are three or more quantities to estimate.

The James-Stein estimator for the first item is 0.779, and this is better (indicated by J in the last column), because it is closer to the true mean, which is unknown to either estimator. The James-Stein estimator is more complicated here, because the number of years of data is different by item. It is necessary to use an iterative procedure to find the best shrinking constant on each item (see Efron and Morris, 1977, for details).

The result is a tremendous improvement in squared error and a large improvement in absolute error. Despite these improvements the James-Stein estimate and the usual estimate are each closer to the true value in 5 cases. Of course, the amount of improvement depends on the observed demands drawn by the simulation as well as the extent to which the true demand rates are similar. James-Stein will always produce a lower expected squared error, but in a specific trial it is possible that the error will be larger.

Table 4-4. James-Stein Simulation Example

Item #	Years Data	Observed Demand	Usual Estimate	True Rate	James-Stein Estimate	Better Estimator
1	4	4	1.000	0.800	0.779	J
2	2	1	0.500	0.800	0.491	U
3	5	3	0.600	0.600	0.565	U
4	4	5	1.250	0.600	0.921	J
5	6	1	0.167	0.400	0.240	J
6	12	4	0.333	0.400	0.340	J
7	3	1	0.333	0.320	0.425	U
8	5	1	0.200	0.320	0.290	J
9	4	1	0.250	0.240	0.353	U
10	8	2	0.250	0.240	0.280	U
Average	Squared	Error	0.063		0.026	-59.1%
Average	Absolute	Error	0.160		0.119	-25.6%

It is instructive to compare the James-Stein estimators for items 6 and 7. The usual estimate of 0.333 is identical for the two items, but the James-Stein estimator for item 6 gives more weight to the observed data because there are 12 years of data as compared to item 7 with only 3 years. This seems very reasonable.

We used another set of data with lower demand rates and more years of history in the simulation. The results in Table 4-5 show less improvement in the errors for James-Stein, but the James-Stein estimator is closer to the true mean for 8 of the 10 items.

Table 4-5. James-Stein Simulation Example - More Years

Item #	Years Data	Observed Demand	Usual Estimate	True Rate	James-Stein Estimate	Better Estimator
1	20	4	0.200	0.200	0.193	U
2	10	3	0.300	0.200	0.273	J
3	25	5	0.200	0.150	0.194	J
4	20	3	0.150	0.150	0.147	U
5	30	4	0.133	0.100	0.132	J
6	60	4	0.067	0.100	0.132	J
7	15	2	0.133	0.080	0.131	J
8	25	0	0.000	0.080	0.007	J
9	20	1	0.050	0.060	0.054	J
10	40	2	0.050	0.060	0.052	J
Average	Squared	Error	0.0024		0.0017	-29.2%
Average	Absolute	Error	0.0370		0.0329	-11.1%

Of course, we have to be a little careful using James-Stein estimators if we believe demand rates change. In Table 4-5 we have one item with 60 years of data (the space station is supposed to last only 30 years). However, as we noted above, there are multiple applications of the items on the station so that the effective years of experience are greater than the elapsed time. We would not recommend the constant demand rate assumptions here for more than a year or two of elapsed time.

4.14 Comparison of Bayes and James-Stein

It is clear that both Bayes and James-Stein estimators shrink the observed estimate toward the overall mean. The James-Stein procedure has one important advantage over Bayes. It can be employed without knowledge of the prior distribution - one not need even believe that such a thing as a prior distribution exists. On the other hand, ignorance has a price in that the James-Stein estimators have a larger expected squared error by an amount proportional to $3/I$, where I is the number of means being estimated. The additional error is therefore negligible when I is greater than 15 or 20, and it is tolerable for I as small as 9. For inventory theory applications where we may be estimating hundreds of means, this difference will be unimportant.

4.15 Demand Prediction Experiment Design

We have performed a variety of demand prediction experiments over the years, including the "objective Bayes" procedure that was field-tested at George Air Force Base, as described in Section 1.9. In this section we want to describe a demand prediction experiment that was performed by Sherbrooke (1987). There are two reasons for presenting this material: (1) a general method is developed for comparing demand prediction techniques; (2) a specific prediction technique is shown to perform best on each of the three aircraft types studied, and thus this prediction technique may be best in other problems as well.

A demand prediction technique in this context is defined to be procedures for estimating both the mean demand and the variance-to-mean ratio of demand over the next year for all first-indenture items on a particular aircraft type at a base. It is important to stress that there is a fundamental difference between demand prediction in a spares context and prediction of economic indicators such as Gross National Product (GNP) or interest rates. In the latter case the forecaster is trying to predict the quantity with minimum error; in our case the logistician is not trying to predict mean demand, but mean demand plus some safety level. That is why it is critical to predict item variance-to-mean ratios as well as means.

The item data consist of quarterly Air Force demand totals and flying hours for each of sixteen quarters, average repair time, and unit cost.

The basic idea of the experiment is to use the first 12 quarters of data and a demand prediction technique to estimate the mean and variance of demand for each item over the next year. Then the optimal availability model of Chapter 2 is used to allocate a fixed investment across the group of items. The result of this is a stock level for every item and a predicted availability over the next year.

The use of an optimal availability model in the assessment of demand prediction techniques is not a common idea, but it is powerful and we recommend it highly. The alternative that is used in most demand prediction studies is to look at error measures by item such as average absolute error, squared error, or percent error. These must be combined across items to get an overall measure, but should these errors just be added up or should they be weighted in some way? There are additional difficulties if a measure such as percent error is used and the true demand is zero. By contrast, the optimal availability model takes care of both the problem of the appropriate item error measure and how to combine across items: (1) the error function is backorders which we know translates to our criterion of interest, availability; (2) the combination of errors across items takes into account that we are likely to have larger backorders on more expensive items.

Then we evaluate the demand prediction technique by seeing how well those stock levels perform over the last four quarters of the 16 quarter data base. Unfortunately, we don't know the day on which each demand occurred, so we are forced to take the quarterly total for each item and randomly assign those demands to days within the quarter. Then we simulate the performance of the supply system and the daily availability by using the demands and the repair times, keeping track of the balance of stock on hand or backorders. The result is an "attained" availability averaged over the year, where the quotes are used to remind us that these availabilities were never actually observed in the real world.

The effect of randomly assigning demand to days within a quarter will tend to improve the availability of every demand prediction technique. Unfortunately there is no good alternative when the data provide only quarterly totals, but any large changes in demand from quarter to quarter during the evaluation year will still affect the prediction techniques since the randomization is done within each quarter subject to the observed quarterly total. Although the performance of each prediction technique will be improved somewhat by the randomization, the relative performance should be affected much less.

The demand prediction experiment is to compare the "attained" availability from a number of demand prediction techniques for a specified total investment in spares. Our primary objective is to find the demand prediction technique that yields the highest "attained" availability. A secondary objective is to find a prediction technique that is properly calibrated in the sense that the predicted and "attained" availabilities are close. The demand prediction experiment is summarized in Figure 4-3.

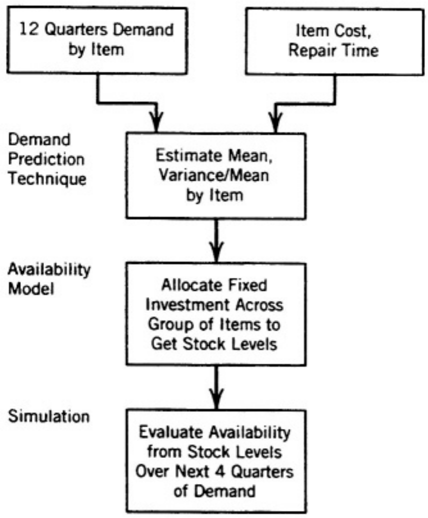


Figure 4-3. **Experimental procedure for demand prediction experiment**

4.16 Demand Prediction Experiment Results

The most significant results from the demand prediction studies are:

1. The best estimator of mean quarterly demand from quarterly data is exponential smoothing:

$$\hat{d} = \omega d_1 + \varpi(1-\varpi)d_2 + \varpi(1-\varpi)^2 d_3 + \ldots \qquad (4.29)$$

where ^ indicates an estimated value, $\hat{d}$ is the estimate of demand for the next quarter, d_1 is the most recent quarter of observed demand, d_2 is the previous quarterly observation, etc. and ω is the smoothing constant. We recommend that with quarterly data the value of ω should be 0.4.

2. The best estimator of variance-to-mean ratio, *V*, is a power function of

the estimated annual mean, $\hat{m} = 4\hat{d}$:

$$\hat{V} = 1 + .14\, \hat{m}^{.5} \tag{4.30}$$

As noted above, these relationships are consistent with our observations that demand is Poisson, but the mean changes or drifts over time. The form of Equation 4.30 guarantees that the variance-to-mean ratio will never be less than that of a Poisson with a constant mean, namely one. Extensive details of the demand prediction experiments are provided in Appendix C. For brevity we mention here that Equations 4.29 and 4.30 performed best on each of the C-5, A-10, and F-16 aircraft systems. On the other hand there is no proof that these relations are optimal – there is merely empirical evidence that they perform well on the systems noted.

Table 4-6 shows that if the best technique for estimating the mean, Equation 4.29, is used in conjunction with a Poisson assumption for the variance-to-mean ratio instead of Equation 4.30, the "attained" availability over the prediction year will be much lower. Also the predicted availability from the Poisson assumption will be far too optimistic. The results are shown for two investment levels and the C-5 aircraft, but similar results are obtained for the other aircraft systems and are discussed in detail in Appendix C.

In other words, suppose that there is $100 million dollars for spares over the next year. We estimate demand over the next quarter using exponential smoothing on the last twelve quarters with a smoothing constant of 0.4. Our estimate of demand over the next year is obtained by multiplying the estimate for the next quarter by four.

Table 4-6. **Demand Prediction on C-5 aircraft**

Budget	Estimator	Variance/Mean Ratio (V)	Attained Availability (%)	Predicted Availability (%)
$80M	Poisson	1	56.6	98.6
	Power Curve	$1+.14m^{.5}$	82.8	77.1
$100M	Poisson	1	62.2	99.9
	Power Curve	$1+.14m^{.5}$	91.3	91.1

In Table 4-6 we compare two alternatives for estimating the variance-to-mean ratio, V. One is to assume that demand has a Poisson distribution with a constant mean (V = 1); the second is to use the power curve relationship. These assumptions lead to different stockage policies. If we assume V = 1, we would predict that over the next year the availability from that stockage policy would be 99.9%, but in fact the attained availability with simulation is only 62.2%. If we assume the power curve relationship to develop our stockage policy, the predicted availability over the next year is only 91.1%, but the attained availability of 91.3% is much higher than under the V = 1 assumption and it is very close to the predicted availability. Note that the simulation makes no assumption about the demand process, because it uses the demands that were actually observed each quarter.

The objective Bayes techniques, discussed in Section 4.10, did almost as well as the variance-to-mean ratio from Equation 4.30. However, Bayes is more complicated, and it cannot be used unless there is demand data, whereas Equation 4.30 can be applied to initial estimates. An interesting finding is that the "system approach" to estimating variance-to-mean ratios given in Equation 4.30 leads to much better estimates of item variance than the past data on the individual items. This is partly because the item means are changing over time, and thus computations of variance around a changing mean are very unstable. It is also because we cannot hope to get better estimates of demand (and variance) just by going back further and using more historical data, since these values change with time.

4.17 Random Failure versus Wear-out Processes

The demand process for some items is not random, but results from wear out. Such items as engine parts, tires, landing gear, gun barrels, batteries, solar arrays are likely to have demand rates that increase around a service life. Another way of looking at these wear-out items is that the probability distribution of time to the next demand does not decrease uniformly like the exponential. Instead there is a peak value to the right of the origin as in distributions such as the gamma, Weibull, or log normal.

The gamma distribution was defined in Equation 4.20 above and shown in Figure 4-2. The exponential is the special case where $a = 1$. Reliability engineers tend to like the Weibull distribution (wei), which is another two-parameter generalization of the exponential:

$$\text{wei}(x) = ab^{-a}x^{a-1}e^{-(x/b)^a} \qquad x \geq 0 \tag{4.31}$$

where the parameters $a, b > 0$. The mean is $(b/a)\Gamma(1/a)$ and

$$\text{variance-to-mean ratio} = b\{2\Gamma(2/a) - (1/a)[\Gamma(1/a)]^2 / \Gamma(1/a)$$

We prefer the gamma distribution to the Weibull, because we can specify a mean and variance-to-mean ratio and compute the parameters of the gamma distribution immediately. By contrast, two nonlinear equations must be solved to determine the Weibull distribution parameters (see Problem 16). But, for a given mean and variance-to-mean ratio the shape of the two distributions is virtually identical as illustrated in Figure 4-4. The Weibull does have an advantage over the gamma distribution. In a simulation where the time to the next failure needs to be drawn probabilistically, the Weibull is easier to sample.

Our interest is not the time until the next demand, but something related to it - the mean and variance-to-mean ratio for the number of demands over the pipeline. Even when the underlying probability distributions are slightly different, the pipeline means and pipeline variance-to-mean ratios will be virtually identical. The important thing is that the variance-to-mean ratio for these state probabilities, the number of demands during some time period T, has a variance-to-mean ratio less than one. As in the case of the negative binomial, we want to be able to take any specified mean and variance-to-mean ratio, estimate the parameters of the best-fitting probability distribution, and then calculate that probability distribution of demand.

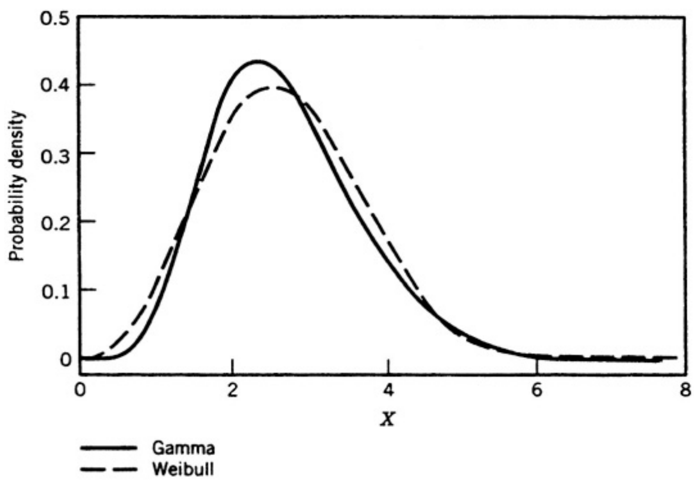


Figure 4-4. **Gamma and Weibull comparison (mean = 2.68; variance/mean = .35).**

It turns out that a particularly convenient choice of distribution is the binomial (bin):

$$\text{bin}(x) = \binom{n}{x} \rho^x (1-\rho)^{n-x} \qquad x = 0, 1, 2 \ldots, n \qquad (4.32)$$

where x can be thought of as the number of successes in n trials when each trial has probability ρ of success. The binomial distribution has mean $n\rho$ and variance-to-mean ratio = $(1 - \rho)$ where ρ is in the range (0, 1) (see Problem 10). Note that this probability distribution is defined for non-negative integers ending at n instead of infinity. When the mean, μ, and variance-to-mean ratio, V, over some time period T are specified, it is easy to solve for the parameters n and ρ since:

$$\mu = n\,p \qquad\qquad V = (1 - p) \tag{4.33}$$

implies that

$$p = (1 - V) \qquad\qquad n = \mu/(1 - V). \tag{4.34}$$

However, we must be a little careful, because the ratio $\mu/(1 - V))$ is not necessarily integral, as required by Equation 4.32. Because of the integer constraint on n, we cannot estimate parameters for arbitrary μ and V; but, we can estimate parameters for an arbitrary μ and a slightly larger V, an arbitrary V and a slightly larger μ, or a μ and V that are each a little larger. (It seems better to overestimate demand probabilities by rounding up rather than down). We prefer the first of the three alternatives which implies that the parameter estimation equations are:

$$n = [\mu/(1 - V) + .99] \qquad\qquad p = \mu/n \tag{4.35}$$

where the bracket notation in the definition for n indicates the integer portion. The addition of .99 inside the bracket insures that if $\mu/(1 - V)$ is integral, that value is used for n; otherwise the value of n is one larger. Thus, ease of parameter estimation is one advantage of the binomial over other state probability distributions with variance-to- mean ratios less than one.

A second advantage of the binomial distribution is that the binomial recursion for x in terms of $x - 1$ is identical with that of the negative binomial (Problem 11). Thus, the same computer program can be used for variance-to-mean ratios less than one or greater than one.

Table 4-7 shows two binomial probability distributions, each with mean one, and variance-to-mean ratios of .5 and .75. We have calculated the mean, variance, and variance-to-mean ratio of backorders as well for comparison with Table 4-2. Note that for any probability distribution, VBO(0)]/EBO(0) equals the variance-to-mean ratio of demand. As s increases VBO(s)]/EBO(s)] increases to a maximum and then declines. An example of binomially distributed demand is provided in Problem 20.

In the next section we consider techniques for testing whether data are Poisson, negative binomial, or binomial. We defer until Chapter 7 a discussion of how more general failure rate distributions can be related to the probability distribution for the number of units of an item that wear out during a period of time.

Table 4-7. Binomial Distributions with Mean = 1

Stock	Var/Mean = .5				Var/Mean = .75			
(s)	bin(s)	EBO(s)	VBO(s)	VBO(s)/EBO(s)	bin(s)	EBO(s)	VBO(s)	VBO(s)/EBO(s)
0	.2500	1.0000	.5000	.5000	.3164	1.0000	.7500	.7500
1	.5000	.2500	.1875	.7500	.4219	.3164	.3335	1.0540
2	.2500	.0000	.0000	.0000	.2109	.0547	.0595	1.0882
3					.0469	.0039	.0039	.9964
4					.0039	.0000	.0000	.0000

4.18 Goodness-of-Fit Tests

In previous sections of this chapter we have assumed that the probability distribution of demand for an item is known, and the problem is to estimate the mean and variance-to-mean ratio. Here we want to consider the inverse problem of testing a set of observed data to see whether it meets the hypothesis of binomial or Poisson demand with a constant mean.

We show how to use the chi-square (χ^2) test for goodness-of-fit in the standard way, which may be familiar to some readers with statistical training. Then we show how chi-square can be used in another much more powerful way to construct an index of dispersion, a test of the variance assumption. Hoel (1962) is the only source we have found for the index of dispersion, and is also an excellent reference for the standard use of chi-square.

Let X be a random variable for the number of demands in each of n time periods of equal length. Let Obs(x) denote the number of these time periods when $X = x$ demands are observed, as shown in the example of Table 4-8. (Note that the sequence in which the demands occur is irrelevant). We want to test whether this data is consistent with the assumption of a Poisson distribution with constant mean, μ. First, the parameters of the probability distribution must be estimated from the data, which for the Poisson is the only the mean. In the example of Table 4-8, we observe 65 periods and a total of 168 demands, so our estimate of μ is $168/65 = 2.58$.

Next we estimate the expected number of periods with x demands, Ex(x), for each value of x. Since we are assuming Poisson demand and have estimated μ, we can calculate $\mathrm{Ex}(x) = np(x|\mu)$, shown in the third column of Table 4-8. For example, $\mathrm{Ex}(0) = (65)p(0|2.58) = 4.9$.

Since the chi-square is only an approximation to the exact distribution, there are some restrictions on its use. The expected number of demands in each cell, $\mathrm{Ex}(x) \geq 5$ and the number of cells, $k \geq 5$. If $k < 5$ then it is best to have the Ex(x) be somewhat greater than 5. Because of the former reason we have combined all cells with 5 or more demands into one cell. It is permissible to combine cells in any way, provided that the method of combination is not influenced by the observed demands.

The chi-square test consists of adding up the quantity in the last column of Table 4-8 over the k distinct values of x, and comparing it with the tabled value for chi-square with k - 1 degrees of freedom that corresponds to the desired significance level. Often a significance level of .05 is chosen which means that if the computed value exceeds the threshold, there is no more than a 5% chance that the data came from a Poisson distribution.

The test is performed for k -1 degrees of freedom, because one parameter, the mean, was estimated from the k cells; more generally, when the chi-square is applied to a problem where several parameters are estimated from the data, the number of degrees of freedom must be reduced by the number of parameters estimated.

When the chi-square test is applied in this standard fashion, we are forced to accept the Poisson assumption because the computed value of 6.66 is less than 11.07, the .05 significance level value for 5 degrees of freedom. Looking at the data in Table 4-6, we note that demands of 8, 15, and 20 are very unlikely from a Poisson with mean 2.58. However, due to the requirement for **Ex(x) ≥ 5,** the test was unable to discriminate between the individual values of x in the last cell where observations were combined.

Table 4-8. Goodness-of-Fit Test

Number of Demands (x)	Time Periods Observed Obs(x)	Time Periods Expected Ex(x)	[(Obs(x) − Ex(x)]^2/Ex(x)
0	10	4.9	5.31
1	14	12.7	.13
2	16	16.4	.01
3	12	14.1	.31
4	7	9.1	.48
5	3		
8	1	7.8	.42
15	1		
20	1		
Mean:	2.58	2.58	Sum = 6.66

Now let's turn to a more powerful version of the chi-square test. It is more powerful than the standard version, because there is a greater chance that we can reject the null hypothesis of Poisson demand with a given set of data.

We begin by supposing that demand in each of the n periods has a binomial distribution where y trials are made in each period resulting in x_i

successes in period *i*. Then $n - x_i$ failures occur in period *i*. These can be arranged in the two-way table shown in Table 4-9.

Table 4-9. **Binomially Distributed Observations**

The mean of each cell in the first row is the expected number of successes

$$\sum_{i=1}^{n} x_i / n = \bar{x} \tag{4.36}$$

and the mean of each cell in the second row is the expected number of failures, $y - \bar{x}$.

Then the chi-square for the contingency table is just

$$\chi^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{\bar{x}} + \sum_{i=1}^{n} \frac{(y - x_i - y + \bar{x})^2}{y - \bar{x}}$$

$$\chi^2 = \left(\frac{1}{\bar{x}} + \frac{1}{y - \bar{x}} \right) \sum_{i=1}^{n} (x_i - \bar{x})^2$$

$$\chi^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{\bar{x}\left(1 - \frac{\bar{x}}{y}\right)} \tag{4.37}$$

where $\bar{x}/y = p$, the probability of success, and the test is called the *binomial index ofdispersion.* The Poisson index of dispersion is obtained by letting $\bar{x}/y$ get very small in Equation 4.37 yielding:

$$\chi^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{\bar{x}} \tag{4.38}$$

When the data of Table 4-8 is substituted into Equation 4.38, the estimate of the mean μ is $\bar{x} = 2.58$, and there are $n - 1 = 64$ degrees of freedom leading to a value of 243.3 for the chi-square. Since the number of degrees of freedom, k, exceeds 30, the chi-square is tested using a normal distribution with mean k and standard deviation $2k$. This value of chi-square is found to be 15.84 standard deviations from the normal distribution mean, which implies that the probability the data came from a Poisson distribution is less than .00005.

The increased power of the second chi-square test is due to the fact that the number of degrees of freedom has increased from 5 to 64. In Table 4-8 the chi-square value for the last cell was quite small, because we had to combine the observations for all demands of five or more. This is precisely the part of the observed data that is most significant in rejecting the Poisson hypothesis in the index of dispersion (Problem 17). However, the Poisson index of dispersion will sometimes accept data as Poisson that is rejected by the standard chi-square test (Problem 18).

Suppose that we want to do a goodness-of-fit test to the negative binomial distribution. It is possible to use the standard approach as in Table 4-8, except that we must subtract two degrees of freedom since the mean and variance-to-mean ratio are estimated from the data. However, there is no negative binomial index of dispersion (Problem 19).

4.19 Summary

We have shown that an exponential probability distribution for the time between demands leads to Poisson (state) probabilities for the number of demands during any specified time period, T. The Poisson distribution was generalized to the negative binomial whose variance exceeds the mean. It has two parameters, allowing us to fit a mean and variance-to-mean ratio separately to observed data. This in turn is useful for two reasons: (1) the number of units of an item in the pipeline in Chapter 3 was treated as if it were a Poisson variable, but the variance usually exceeds the mean, often substantially; (2) when the demand rate changes or drifts with time, the variance will exceed the mean.

We discussed Bayesian analysis and suggested that in many problems good agreement between different analysts can be obtained for the prior distribution. Sometimes "objective" Bayes procedures can be applied, as in the field test of the Base Stockage Model at George Air Force Base. Empirical Bayes techniques can be applied to some Bayes problems if the sample size is very large. We discussed a standard Bayesian procedure to combine demand data with initial estimates by item. This requires the analyst to estimate the relative "information" in the initial estimate and demand over a period of time.

An alternative approach that requires no prior distribution is the use of James-Stein estimators. Bayes and James-Stein proceed from different philosophical premises, but they tend to produce very similar estimates.

Demand prediction experience was discussed and it was concluded that exponential smoothing for the estimation of mean demand and a power formula based on the estimated mean were the best techniques for each of three different aircraft systems. The objective Bayes technique did almost as well, but was not recommended because of its additional complexity and its inapplicability to problems where there are only initial estimates. The topic of demand prediction is given more detailed attention in Appendix C.

Then we showed that items whose failure is dominated by wear-out phenomena can be modeled with a binomial distribution which has a variance-to-mean ratio less than one. The binomial distribution is easy to use, and has the same recursion formula as the negative binomial.

Finally, we discuss the problem of testing observed data to determine whether it may be binomial, Poisson, or something else. In addition to the usual chi-square goodness-of-fit test we derive expressions for the binomial and Poisson indices of dispersion. These tend to be much more powerful tests with a greater capability of rejecting the null hypothesis because the number of degrees of freedom is greater.

4.20 Problems

1. Combinations

Suppose that there are four possible outcomes a_1, a_2, b_1, b_2, with probabilities .1, .2, .3, and .4 respectively. The probability of a particular sequence of the four distinct outcomes, for example, $a_1a_2b_1b_2$, is just the product of the probabilities. Show that the probability of getting each outcome once in a series of 4 trials is .0576. Write out all the possible sequences and show that there are 4! = 24.

2. Suppose that outcomes a_1 and a_2 are now indistinguishable and that b_1, and b_2 are now indistinguishable. Use the combinatorial formula in Section 4.3 to show that the probability of 2 *a*'s and 2 *b*'s is now .2646. Write out all the possible distinguishable sequences and show that there are 6.

3. Using Equations 2.1 and 2.2, verify the mean and variance-to-mean ratio of the negative binomial as shown in Section 4.3. Also demonstrate that the formulas for a and b given in Equation 4.8 are correct.

4. Derive a recursion formula for neg(x + 1) in terms of neg(x) and multiplicative terms for the negative binomial of Equation 4.5.

5. Show that as the variance-to-mean ratio of the negative binomial, 1/(1 - b), approaches one, the negative binomial probability of zero, neg(0), of Equation 4.5 converges to the Poisson probability of zero, $p(0)$, for any fixed mean. (Hint: Hold the mean, $ab(1 - b)$, constant and show that as b approaches zero the logarithm of neg(0)/p(0) approaches 0. Use the power series expansion for log(1 - b) following Equation 2.19.)

6. Verify the recursion formula for the second moment of the backorder function as given by Equation 4.15. Use this recursion formula to compute VBO(s) for s = 0, 1, 2, 3 when the mean is one and demand is Poisson. The answers are given in Table 4-2.

7. Verify the values of EBO(1) and VBO(1) in Table 4-2 for the negative binomial and in Table 4-4 for the binomial.

8. Since the numerical, multi-indenture example of Section 4.5 gave an answer outside the 95% confidence interval from the simulation, why not simulate instead of using an analytic model?

9. Show that the posterior distribution $h(\mu|x)$ in Equation 4.20 is a gamma distribution where the prior distribution parameters a and b are replaced by $(a + x)$ and $(b/b + 1)$ respectively.

10. Using Equations 2.1 and 2.2, verify the mean and variance-to-mean ratio of the binomial distribution as shown in Section 4.17.

11. Show that the recursion formula for bin(x + 1) in terms of bin(x) and multiplicative terms for the binomial of Equation 4.32 is the same as that for the negative binomial derived in Problem 4.

12. Show that the Poisson distribution has the "independent increments" property of Section 4.3 for any non-negative values of y (discrete), μ_1, and μ_2:

$$p(y|\mu_1 + \mu_2) = \sum_{x=0}^{y} p(x|\mu_1)p(y - x|\mu_2)$$

This is also called a *convolution,* and it is easy to evaluate.

13. Show that the binomial distribution of Section 4.17 has the "independent increments" property of Section 4.3 for any non-negative, discrete values of y, n_1, and n_2 and fixed p in the range $0 < p < 1$ (equivalent to a fixed variance-to-mean ratio, V):

$$\text{bin}(y|n_1 + n_2, p) = \sum_{x=0}^{y} \text{bin}(x|n_1, p)\text{bin}(y - x|n_2, p)$$

Hint: Establish the identity:

$$\binom{n_1+n_2}{y}=\binom{n_1}{0}\binom{n_2}{y}+\binom{n_1}{1}\binom{n_2}{y-1}\cdots\binom{n_1}{y}\binom{n_2}{0}$$

by expanding each bracket of

$$(a+b)^{n_1}(a+b)^{n_2}=(a+b)^{n_1+n_2}$$

for arbitrary a, $b > 0$ and setting coefficients of $a^{n_1+n_2-y}b^y$ for each value of y equal on both sides of the equation. Note that the first identity establishes that the hypergeometric is a probability distribution, i.e. for fixed n_1, n_2, and y the individual probability terms on the right after division by the left-hand side sum to one.

14. Show that the negative binomial distribution (neg) has the "independent increments" property of Section 4.3 for any y, m_1, and m_2 for a fixed variance-to-mean ratio, V:

$$\text{neg}(y|m_1+m_2,V)=\sum_{x=0}^{y}\text{neg}(x|m_1,V)\text{neg}(y-x|m_2,V)$$

15. If a, b are the gamma distribution parameters, show that the denominator of Equation 4.20 has a negative binomial probability distribution with parameters $a'=a, b'=b/(b+1)$.

16. Figure 4-4 was generated by selecting Weibull parameters $a = 3$, $b = 3$. This resulted in a mean of 2.6802 and a variance-to-mean ratio of .3533. Then these values were used to estimate the parameters of the gamma distribution.

Suppose that it is desired to have a Weibull distribution with mean = 4 and variance-to-mean ratio = 1. Using a table of gamma functions (Table 4-10), make two iterations to attempt estimation of the Weibull parameters. Compare this with the difficulty of estimating the gamma distribution parameters.

17. Using the data of Table 4-8, show that the chi-square value for the Poisson index of dispersion from Equation 4.38 equals 243.3. Note which observations make the largest contribution to this value of chi-square.

18. The Poisson index of dispersion is usually a more powerful test than the standard chi-square test because there are more degrees of freedom. However, it will accept the Poisson hypothesis if the overall variance-to-mean ratio is about one, regardless of the individual observations which may be highly non-Poisson. Construct an example that is accepted by the index of dispersion and rejected by the standard chi-square test. Hint: Suppose that there are 100 observations, of which 50 are 0's and 50 are 2's so that the

mean and variance-to-mean ratio are each one. The Poisson probabilities are given in Table 4-2.

19. Show that there is no negative binomial index of dispersion, because the derivation of Equation 4.37 requires that observations can be characterized as in Table 4-9. Note that the Poisson index of dispersion can be defined even though the number of trials y is unknown, because $\bar{x}/y$ is a constant.

20. Let M be the average demand rate for an item per location on an aircraft per flying hour. Suppose the sortie length is t hours. Normally the expected demands per location per sortie, Mt, is much smaller than one. However, suppose that because of aerial refueling or other reasons the sortie duration becomes very long and the product, Mt, is near one. Since an item in a particular location can fail once or not at all in a sortie, the Poisson demand assumption is not appropriate. Find a general formula to estimate the correction factor necessary (due to the fact that after the first demand at a location during the sortie, there can be no further demands), and show that when $Mt = 1$ the mean should be multiplied by .63. Find the general formula for the impact on the variance-to-mean ratio, and show that it is about .37 when $Mt = 1$.

Table 4-10. Values of $\Gamma(x) = (x - 1)!$

x	0	1	2	3	4	5	6	7	8	9
1.0	1.0000	.9943	.9888	.9835	.9784	.9735	.9687	.9642	.9597	.9555
.1	.9514	.9474	.9436	.9399	.9364	.9330	.9298	.9267	.9237	.9209
.2	.9182	.9156	.9131	.9108	.9085	.9064	.9044	.9025	.9007	.8990
.3	.8975	.8960	.8946	.8934	.8922	.8912	.8902	.8893	.8885	.8879
.4	.8873	.8868	.8864	.8860	.8858	.8857	.8856	.8856	.8857	.8859
.5	.8862	.8866	.8870	.8876	.8882	.8889	.8896	.8905	.8914	.8924
.6	.8935	.8947	.8959	.8972	.8986	.9001	.9017	.9033	.9050	.9068
.7	.9086	.9106	.9126	.9147	.9168	.9191	.9214	.9238	.9262	.9288
.8	.9314	.9341	.9368	.9397	.9426	.9456	.9487	.9518	.9551	.9584
.9	.9618	.9652	.9688	.9724	.9761	.9799	.9837	.9877	.9917	.9958

Chapter 5

VARI-METRIC: A MULTI-ECHELON, MULTI-INDENTURE MODEL

There is no royal road to geometry.

–Euclid

5.1 Chapter Overview

In Chapter 3 we presented the METRIC theory for the multi-echelon problem and demonstrated the optimization procedure. METRIC was the first practical application of multi-echelon inventory techniques, and it has formed the theoretical foundation for a number of models used by the military services of the United States and its allies as well as a number of private companies.

When we developed METRIC, we knew that it understated base backorders. In most cases the error was not large, and the simplicity seemed more important than the lack of precision.

Subsequently other investigators derived exact solutions, usually at the expense of more restrictive assumptions; for example, Simon (1971) assumed constant instead of arbitrary resupply times and Poisson instead of compound Poisson demand. Kruse (1979) developed an exact solution for the multi-echelon problem. However, both Kruse and Simon's models require substantial computer time, which is an important consideration, particularly when modeling many echelons or indentures.

Slay (1984) devised an improvement to METRIC that he called VARI-METRIC, and Graves (1985) more recently published a simple derivation of this approximation. The advantage of the VARI-METRIC technique is that it is much easier to compute than previous improvements. Though Graves found it necessary to assume constant resupply times, he retained the compound Poisson demand assumption. Graves showed that in 11% of cases, the METRIC stock levels differ by at least one unit from the optimal results; the VARI-METRIC levels differ in only 1% of cases.

We demonstrated in the previous chapter that the METRIC approach as applied to the multi-indenture problem can lead to substantial underestimation of backorders. It was shown that by taking into account not only the mean pipeline values, but the variance as well, it is possible to obtain a much better estimate of backorders. This is an example of the VARI-METRIC technique, for which we present in this chapter the combined multi-echelon, multi-indenture theory as published in Sherbrooke (1986).

Before dealing with the combined problem, we derive the mean and variance for the number of units in the pipeline in the multi-echelon case, which requires the proof of a statistical theorem. Then in Section 5.3 we describe the "physics" of the combined multi-echelon, multi-indenture problem. The demand at the base for the first-indenture line-replacable unit (LRU), in combination with the various item parameters such as repair probabilities, allows us to compute the demand for the LRU at depot and for the second-indenture shop-replaceable unit (SRU) demand at base and depot.

Then starting with depot demand for the SRU, we compute the depot pipeline for SRUs and the delay at the depot for LRU repair and at the bases for SRU resupply. Eventually the delay at the base for the LRU is calculated and the expected backorders, from which we can compute availability.

We discuss the calculation of optimal policies and show how to generalize the assumption of constant resupply times that do not vary by base. The Poisson demand assumption is generalized to the cases where (1) the mean changes over time; (2) failures are due to wear-out rather than to random causes.

In Section 5.12 we address the topic of commonality, where a specific SRU may be used on more than one LRU. Clearly commonality is desirable, because the total stock required is less in that case. Then we show how we can determine reorder points and order quantities for items at each echelon. This is particularly important at the depot and for items that are inexpensive or have little or no chance of being repaired.

We discuss how to take into account that all first-indenture items are not equally critical to the operation of the aircraft. Then we show how the availability due to spares can be modified to include the degradation due to remove-and-replace maintenance as well as to periodic maintenance.

The theory in this chapter applies to the case where repair/resupply can be initiated at any time (continuously) and where cannibalization is not performed to consolidate "holes" on the fewest end items. The case of periodic resupply is addressed in Chapters 6, 7 and cannibalization is covered in Chapter 8. Another case of interest is where there is no resupply; the spares may be for a flyaway kit or a war reserve spares kit that must support independent operations for a period of time. In Section 9.5 we show that this problem is just a special case of VARI-METRIC.

5.2 Mathematical Preliminary: Multi-Echelon Theory

In Sections 4.4 and 4.6 we derived equations for the mean and variance of the number of units in the pipeline for the multi-indenture case. Here we need to develop the equations for the multi-echelon case before considering the combined problem.

If we assume that the depot to base order and ship time, O, is a constant that does not vary by base, the resupply delay to any base at an arbitrary time, t, is a function of the status at depot at time t - O. We assume that the depot fills orders on a first-come, first-served basis. Then the distribution of outstanding orders to base j, conditional on the total number of outstanding orders, has a binomial distribution. We will use this fact repeatedly below when we calculate the VARI-METRIC multi-echelon equations. The reader who is not interested in the mathematical details may prefer to skip this section and accept the results embodied in Equations 5.8 and 5.11. The significance of these multi-echelon equations is that the pipeline variance always exceeds the mean unless the depot stock level is zero. This is analogous to our multi-indenture results in Section 4.6 where we found that the LRU pipeline variance always exceeds the mean unless each SRU stock level is zero.

The number of units in the pipeline for base j, x_j, is related to the number of units in repair at the depot, x_0. The expected number of units in the pipeline can be written in terms of the conditional expectation in a manner similar to Equation 4.23

$$E[X_j] = \sum_{x_0=0}^{\infty} E[X_j | X_0]\Pr\{X_0 = x_0\} \tag{5.1}$$

The variance is more complicated and requires proof so we state the following:

LEMMA: Let X_j and X_0 be two variables that are not independent of each other. Then the variance of X_j can be computed as

$$\text{Var}[X_j] = E[\text{Var}[X_j|X_0]] + \text{Var}[E[X_j|X_0]] \tag{5.2}$$

PROOF: The terms on the right-hand side are interpreted in the proof below, which follows Parzen (1962). The variance was defined in Equation 2.2:

$$Var[X_j] = E[(X_j - E[X_j])^2] = E[E[(X_j - E[X_j])^2|X_0]] \tag{5.3}$$

The last equality is just Equation 5.1. But for any random variable Y and constant a

$$E[(Y-a)^2] = E[(Y - E[Y]\}^{2]} + (E[Y]-a)^2 \tag{5.4}$$

Using this relationship, the inner bracket on the right-hand side of Equation 5.3 for any fixed $X_0 = x_0$ becomes:

$$E[(X_j - E[X_j])^2|x_0]$$

$$= E[(X_j - E[X_j|x_0])^2|x_0] + (E[X_j|x_0] - E[X_j])^2 \tag{5.5}$$

Upon taking expectations of both sides in Equation 5.5 with respect to x_0, we obtain

$$\text{Var}[X_j] = \sum_{x_0=0}^{\infty} \text{Var}[X_j|X_0]\Pr\{X_0 = x_0\}$$
$$+ \sum_{x_0=0}^{\infty} (E[X_j|X_0] - E[X_j])^2 \Pr\{X_0 = x_0\}$$

which is another way of writing Equation 5.2, proving the lemma.

Now we want to evaluate Equations 5.1 and 5.2 for the multi-echelon problem where the LRU stock levels are s_j at base j and s_0 at the depot. To evaluate the inner bracket in Equation 5.1 we note that if $x_0 \le s_0$ there are no depot backorders delaying a resupply to the base. Thus, the average number of units in the base pipeline is just the average number in resupply:

$$E[X_j|x_0] = m_j O \qquad x_0 \leq s_0 \tag{5.6}$$

where m_j is the average annual demand by base j on depot for resupply (we generalize to allow some base repair later) and O is the order-and-ship time to any base.

When $x_0 > s_0$ there are $x_0 - s_0$ depot backorders, of which an average fraction m_j / m_0 are from base j. Thus, the expected number of units in the pipeline to base j is:

$$E[X_j|x_0] = m_j O + m_j (x_0 - s_0)/m_0 \qquad x_0 > s_0 \tag{5.7}$$

and taking the expectation with respect to x_0 in Equations 5.6 and 5.7 shows that Equation 5.1 equals:

$$E[X_j] = m_j O + m_j \mathrm{EBO}(s_0)/m_0 \tag{5.8}$$

Now consider Equation 5.2. To evaluate the second term on the right we need the variance of the quantities on the left-hand side of Equations 5.6 and 5.7 for a specified x_0. Since the variance of a constant, $m_j O$, is zero, and the variance of a constant times a variable is the square of the constant times the variance of the variable we have:

$$\mathrm{Var}[E(X_j|X_0)] = m_j^2 VBO(s_0)/m_0^2 \tag{5.9}$$

The inner bracket of the first term on the right of Equation 5.2 is very similar to Equations 5.6 and 5.7, where we utilize the fact that there is a binomial distribution of the $x_0 - s_0$ depot backorders to base j.

$$\mathrm{Var}[X_j|x_0] = \begin{cases} m_j O & x_0 \leq s_0 \\ m_j O + (m_j/m_0)(1 - m_j/m_0)(x_0 - s_0) & x_0 > s_0 \end{cases} \tag{5.10}$$

When Equation 5.9 is added to the expectation of Equation 5.10 with respect to x_0, Equation 5.2 becomes:

$$\text{Var}[X_j] = m_j O + (m_j / m_0)(1 - m_j / m_0)\text{EBO}(s_0) + m_j^2 \text{VBO}(s_0) / m_0^2 \tag{5.11}$$

Equations 5.8 and 5.11 will be used repeatedly in Sections 5.5 to 5.7 below, when we need to compute multi-echelon pipelines. Note that when m_j equals m_0, as in the multi-indenture problem, Equation 5.11 reduces to Equation 4.16 for a single SRU with the proper notational changes.

5.3 Definitions

The combined multi-indenture, multi-echelon process begins when an LRU fails and is brought into base supply. If base supply has a spare LRU, it is issued; otherwise a base LRU backorder is incurred. The failed LRU has a probability of being repaired at the base; otherwise, if the repair is too complex, the LRU is sent to the depot for repair and a resupply request for the LRU is placed on depot.

If the LRU is repaired at the base, we assume that one and only one SRU will be found to have failed. If a spare SRU is available, it is put on the LRU and the LRU repair is completed. The SRU has a probability of being repaired at the base; otherwise the SRU is sent to the depot and a resupply request for the SRU is made on the depot.

When an LRU repair/resupply is completed, a backorder is satisfied if there are any outstanding; otherwise, the LRU stock on hand is increased by one. If the LRU is not repaired at the base, a similar process for SRU repair occurs at the depot. We assume that all SRUs can be repaired at the depot. Later this assumption is relaxed.

Next, define the following constants with the convention that i denotes an SRU item, 1, 2, . . . , I (0 = LRU) and j denotes a base, 1, 2, . . . , J (0 = depot):

m_{ij} = average annual demand for SRU i at base j

T_{ij} = average repair time (in years) for SRU i at base j

r_{ij} = probability that a failure of SRU i at base j can be repaired at that base

q_{ij} = conditional probability that an LRU being repaired at base j will result in a fault isolation to SRU i where $\sum_{i=1}^{I} q_{ij} = 1$

O_i = constant order-and-ship time from depot to any base of SRU i *if the depot has stock on hand.* Note that this time does not vary by base. This assumption is relaxed in Section 5.10.

s_{ij} = stock level for SRU i at base j

As before, a stochastic variable is needed for the pipeline: X_{ij} = number of units of SRU i at base j that are in repair or resupply at a random point in time

It will be useful to modify the notation for expected backorders slightly to include a second argument, Var, for the variance of the state probabilities. When the state probabilities are Poisson, and hence the variance equals the mean, the second argument will be suppressed. $\text{EBO}(s|\mu, \text{Var})$ = expected backorders when the stock level is s and the state probabilities have mean μ and variance Var. The notation for the variance in backorders will be expanded similarly to $\text{VBO}(s|\mu, \text{Var})$.

5.4 Demand Rates

We assume that demand is Poisson until Section 5.11, where that assumption is generalized. It is necessary to develop equations for SRU demand rates and LRU depot demand rates as a function of the base LRU demand rates, m_{0j} for $j > 0$. The sequence of these derivations is shown by the arrows in Figure 5-1a.

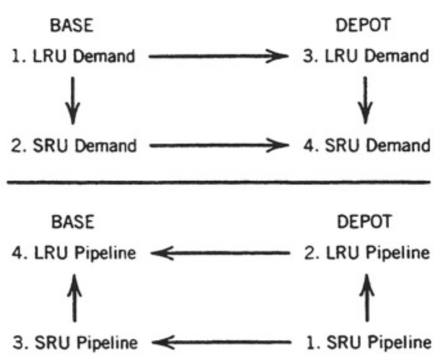


Figure 5-1. **a. Base-depot demand calculation sequence. b. Backorder calculation sequence**

The average annual demand for SRU i at base j is the average annual demand for the LRU at the base times the probability that the LRU is repaired there times the probability that the LRU repair results in a demand for SRU i:

$$m_{ij} = m_{0j} r_{0j} q_{ij} \qquad i, j > 0 \tag{5.12}$$

The average annual demand for the LRU at the depot is the sum of LRU demands at the bases that result in resupply of LRUs from the depot:

$$m_{00} = \sum_{j=1}^{J} m_{0j}(1 - r_{0j}) \tag{5.13}$$

The average annual demand at the depot for SRU i is the resupply demand summed across the J bases (average annual demand for SRU i at base j multiplied by the probability that SRU i is not repairable at the base) plus the average annual SRU demand resulting from LRU repairs at the depot. (Note that the standard MOD-METRIC model of Section 4.4 ignores this second source of SRU demand)[1]. Therefore,

$$m_{i0} = \sum_{j=1}^{J} m_{ij}(1 - r_{ij}) + m_{00}q_{i0} \qquad i > 0 \tag{5.14}$$

The diagram in Figure 5-1b shows the sequence of pipeline and backorder calculations required to compute the expected base backorders for the LRU. This is exactly the reverse of the sequence used to calculate the demand rates above. Our focus is one item group (an LRU and its second-indenture SRUs).

5.5 Mean and Variance for the Number of LRUs in Depot Repair

The fraction of depot demand for SRU i due to depot LRU repairs is

$$f_{i0} = m_{00}q_{i0} / m_{i0} \qquad i > 0 \tag{5.15}$$

The number of LRUs in depot repair consists of two groups: (1) the number of LRUs in the depot repair pipeline when there are no delays for SRUs; (2) LRUs delayed in repair because of not having the required SRUs on hand at depot. For the same reasons underlying Equation 4.12, these two groups are independent. A particular SRU backorder has a probability f_{i0} that it is delaying an LRU repair at depot and a probability $1 - f_{i0}$ that it is delaying a resupply to some base. For any total number of backorders on

[1] MOD-METRIC ignored demand for SRUs caused by depot repair of LRUs, because it would have required a more complex optimization procedure. We use a different optimization procedure, which allows us to include this important source of demand.

SRU i at depot, the probability distribution for the number of them that are delaying LRU repair has a binomial distribution. Thus the mean and variance for the number of LRUs in depot repair are from Equations 5.8 and 5.11:

$$E[X_{00}] = m_{00}T_{00} + \sum_{i=1}^{I} f_{i0}\text{EBO}(s_{i0}|m_{i0}T_{i0}) \tag{5.16}$$

$$\text{Var}[X_{00}] = m_{00}T_{00} + \sum_{i=1}^{I} f_{i0}(1-f_{i0})\text{EBO}(s_{i0}|m_{i0}T_{i0})$$

$$+ \sum_{i=1}^{I} f_{i0}^2\text{VBO}(s_{i0}|m_{i0}T_{i0}) \tag{5.17}$$

5.6 Mean and Variance for the Number of SRUs in Base Repair or Resupply

Analogous to the definition of f_{i0}, let f_{ij} be the fraction of all demand at the depot for SRU i that is being resupplied to base j:

$$f_{ij} = m_{ij}(1-r_{ij})/m_{i0} \qquad i,j>0 \tag{5.18}$$

Equations 5.14, 5.15, and 5.18 imply that

$$\sum_{j=0}^{J} f_{ij} = 1 \qquad i>0$$

The number of SRU i in repair or resupply to base j is the number of units in the pipeline plus the delay due to the fact that the depot does not always have the SRU on hand. The fraction of the latter delay affecting base j has a binomial distribution again (since O_i is assumed to be a constant that does not vary with the base j). Equations 5.8 and 5.11 yield the mean and variance for the number of SRU i in repair at base j or being resupplied to base j:

$$E[X_{ij}] = m_{ij}[(1-r_{ij})O_i + r_{ij}T_{ij}] + f_{ij}\text{EBO}(s_{i0}|m_{i0}T_{i0}) \qquad i,j>0 \tag{5.19}$$

$$\mathrm{Var}[X_{ij}] = m_{ij}[(1-r_{ij})O_i + r_{ij}T_{ij}] + f_{ij}(1 - f_{ij})\mathrm{EBO}(s_{i0}|m_{i0}T_{i0})$$

$$+ f_{ij}^2\mathrm{VBO}(s_{i0}|m_{i0}T_{i0}) \qquad i,j>0 \qquad (5.20)$$

5.7 Mean and Variance for the Number of LRUs in Base Repair or Resupply

Let f_{0j} be the fraction of LRU demand that came from base j:

$$f_{0j} = m_{0j}(1-r_{0j})/m_{00} \qquad j>0 \qquad (5.21)$$

Note that because of Equation (5.13)

$$\sum_{j=0}^{J} f_{0j} = 1$$

The number of LRUs in repair/resupply at base j may be divided into three groups: 1) the number of units of the LRU in the pipeline when there are no backorders; 2) the LRU resupply delay due to LRU backorders at the depot; 3) the base LRU repair delay due to SRU backorders at base. Since each LRU failure is assumed to be due to the failure of one SRU only, the mean and variance of the third group is just the sum of means and variances for the SRUs as in Equations 4.12 and 4.16:

$$E[X_{0j}] = m_{0j}[(1-r_{0j})O_0 + r_{0j}T_{0j}] + f_{0j}\mathrm{EBO}(s_{00}|E[X_{00}],\mathrm{Var}[X_{00}])$$

$$+ \sum_{i=1}^{I}\mathrm{EBO}(s_{ij}|E[X_{ij}],\mathrm{Var}[X_{ij}]) \qquad j>0 \qquad (5.22)$$

$$\mathrm{Var}[X_{0j}] = m_{0j}[(1-r_{0j})O_0 + r_{0j}T_{0j}]$$

$$+ f_{0j}(1 - f_{0j})\mathrm{EBO}(s_{00}|E[X_{00}],\mathrm{Var}[X_{00}])$$

$$+ f_{0j}^2\mathrm{VBO}(s_{00}|E[X_{00}],\mathrm{Var}[X_{00}])$$

$$+ \sum_{i=1}^{I}\mathrm{VBO}(s_{ij}|E[X_{ij}],\mathrm{Var}[X_{ij}]) \qquad j>0 \qquad (5.23)$$

Under the assumption of Poisson demand the variance of the number of SRUs in depot repair was equal to the mean number in depot repair. But unless the stock level for SRUs at the depot is zero, the variance of the number of LRUs in depot repair in Equation 5.17 is larger than the mean in Equation 5.16. Similarly unless the stock level for SRUs at the depot is zero, the variance of the number of SRUs in base repair or resupply in Equation 5.20 is larger than the mean in Equation 5.19. Finally the variance of the number of LRUs in base repair or resupply in Equation 5.23 is even larger when compared to the mean number in repair or resupply in Equation 5.22. This is because the state probabilities on the right-hand side of these equations have variances that exceed their means unless all stock levels at the depot and all SRU stock levels at bases equal zero.

5.8 Availability

The availability at base j due to expected backorders on the LRU and its SRUs is given by

$$A_j = 100\{1 - \mathrm{EBO}(s_{0j}|E[X_{0j}],\mathrm{Var}[X_{0j}])/(N_j Z_0)\}^{Z_0} \quad j>0 \quad (5.24)$$

where N_j is the number of aircraft at base j and Z_0 is the number of applications of the LRU (QPA) on the aircraft. The right-hand side of Equation 5.24 must be multiplied by a similar bracketed term for every LRU family to obtain the overall availability at base j, and the availabilities at each base j must be multiplied by the number of aircraft using Equation 3.3 to obtain the overall system availability. As before, if $\mathrm{EBO}(s_{0j}) > N_j Z_0$ for any item, $A_j = 0$.

The minimization of the sum of base LRU backorders is equivalent to maximization of availability only if the number of aircraft is the same at each base. Otherwise it is necessary to weight the base LRU backorders by base in Equation 5.24.

5.9 Optimization

We have shown how to calculate the expected base backorders as a function of depot and base stock levels for an LRU and its second-indenture SRUs. The optimization procedure consists of augmenting these levels and retaining those solutions that lie on the convex hull. A similar optimal function for the sum of base backorders versus cost must be computed for each LRU family (i.e. the LRU and its SRUs), and then marginal analysis is used to combine across LRU families to obtain the system solution.

5.10 Generalization of the Resupply Time Assumptions

In the previous sections we assumed that the depot to base resupply time, O_i, is a constant that may vary by item but not by base. As a practical matter we are more likely to want to model variation by base (e.g. U.S. vs. overseas bases) rather than by item. Also, it would be nice to generalize the assumption of a constant time to an arbitrary probability distribution of resupply time.

These generalizations cause no theoretical difficulty in the Calculation for the mean and variance of the number of units in the resupply pipeline (e.g. the first part of Equation 5.19, because Palm's theorem is applicable. The problem arises in the allocation of depot backorders for a particular item i to the several bases. We have no guarantee that the binomial distribution is appropriate, because the status at each base at an arbitrary time, t, relates to the depot status at different earlier times, $t - O_{ij}$.

However, it is likely that the binomial is not an unreasonable approximation in the more general case. Equations 5.19 and 5.22 for the expected values should still be correct; Equations 5.20 and 5.23 for the variances will be wrong if the distribution is not binomial, but the error should be slight, particularly if the depot backorders are not too large. We would expect that the approximation would tend to be better as the number of bases increases as well. Finally, it is important to remember that VARI-METRIC is an approximate model anyway. If the order and ship times do vary by base, that should be reflected in the model input because the pipelines are computed correctly even if the binomial allocation of backorders to bases is only approximately correct. (see Problem 7).

5.11 Generalization of the Poisson Demand Assumption

In Chapter 4 we noted that real world data suggest that there are demand processes other than Poisson with a constant mean: (1) random failures from a Poisson process with a changing mean that lead to a variance-to-mean ratio greater than one; (2) failures due to wear out such that the variance-to-mean ratio is less than one. In Tables 4.2 and 4.5 we provided examples for the mean and variance of backorders when the state probabilities are negative binomial or binomial respectively.

Let's consider Equations 5.16 and 5.17 and see what changes are required to approximately model demand processes with variance-to-mean ratios that are not one. The first term on the right-hand side of Equation 5.16 is the expected number of LRUs in the repair pipeline and is unaffected. The expected backorders in the second term for each SRU i are computed for a mean demand $m_{i0}T_{i0}$ over a time period T_{i0}.

We start with the case of variance-to-mean ratios greater than one arising from a Poisson process with a changing mean. Suppose that we believe an estimation equation such as Equation 4.30 is appropriate for estimating the variance-to-mean ratio of annual demand from an estimate of the annual mean. We substitute m_{i0} for $\hat{m}$ in Equation 4.30 to obtain an estimate of the variance-to-mean ratio, V. Then, the expected backorders on the right-hand side of Equation 5.16 are computed using a negative binomial distribution with a mean of $m_{i0}T_{i0}$ and the variance-to-mean ratio, V, just computed. This must be done for each SRU i, and we will obtain different values of V usually. Similar considerations apply to Equation 5.17 except that the first term on the right-hand side, $m_{00}T_{00}$, must be multiplied by the variance-to-mean V for the LRU, obtained by substituting into Equation (4.30) using m_{00} as the average annual demand.

Now we turn to variance-to-mean ratios less than one arising from items with wear-out characteristics. If we assume that SRU i failures are primarily due to wear out, then we can evaluate Equations 5.16 and 5.17 using binomial state probabilities. Here we have no estimation equation for variance-to-mean as a function of the mean, so we would typically assume that the variance-to-mean ratio is independent of the mean value. However, this does require the user to make an estimate of variance-to-mean ratio. We will return to this topic in Chapter 7. Of course, an alternative that can be used is to assume that demand for these wear-out items is Poisson. This overstates variance, but that is less severe than understating the variance of Poisson demand processes with changing means.

5.12 Common Items

Suppose that a specific SRU is used on two (or more) different LRUs. The theory above has ignored this possibility, with the result that the total stock for the SRU, computed as if the SRU were unique in the two locations, would be unnecessarily large (or the backorders would be overstated). Obviously commonality is desirable.

The mathematical modifications necessary to model commonality are straightforward. For example, consider the modifications to Equations 5.16 and 5.17 for the mean and variance of the number of units of the LRU in depot repair. m_{i0} must now be the total depot demand for SRU i from all LRUs on which it is common. Similarly, f_{i0} is now the fraction of total depot demand for SRU i that arises from depot repair of a particular LRU parent. Only a portion of the backorders for the common SRU at depot contribute to the pipeline delay at depot for a particular LRU parent. As before, this portion has a binomial probability distribution, so that Equations 5.16 and 5.17 are still valid with the revised definitions of m_{i0} and f_{i0}. Of course, the effect of the SRU must be computed for all LRU parents, and comparable changes made in the Equations 5.22 and 5.23 for the mean and variance of the number of units of each LRU parent in base repair.

The SRU can be common to many LRUs, and there can be more indentures and echelons. This does complicate the computer programs substantially (particularly because sub-SRUs could be common to several SRUs some of which in turn are common to multiple LRUs), but the basic logic is the same. The computer programs have to be even more sophisticated to take into account that a common item does not always appear at the same indenture level.

5.13 Consumable and Partially Repairable Items

We have considered only repairable items. As explained in Chapter 1, this simplifies the stockage problem in that it is appropriate to use an (s - 1, s) inventory policy. We buy a stock level s and the quantity to order or repair is one. Now we must find a way to compute the reorder point, R, and order quantity, Q, for lower-indenture items which may have large demand and low cost.

The mathematical problem is that it is more difficult to find the optimum values of two decision variables, R and Q, on each item rather than just the stock level. Muckstadt (1982) suggested that a simple, approximate approach would be to use the (s - 1, s) inventory theory of the repairable item models to calculate the optimal stock level for each item at each location and the expected backorders. Using the expected backorders for a particular item and location as a constraint, it is easy to calculate on a single item basis the optimal Q and R.

It turns out that when the resulting Q*'s* are greater than one, it is most likely to be at the depot where demand is largest and for inexpensive items that are seldom repairable. The latter is due to the fact that the average depot pipeline time is the procurement lead time rather than the depot repair time for such items, and the procurement lead time is typically longer. Also, the external order cost for a procurement is typically larger than the internal order cost between echelons of the supply system.

Of course, the economic order quantity, Q, may be greater than one for other types of items and at other echelons. It should be noted that in each case the reorder point and order quantity for resupply from the next higher echelon is being computed. The demand that is repaired locally at the site does not enter into the calculation of Q and R.

To perform this calculation we need the probability distribution of lead time demand. Since the variance of lead time demand can exceed the mean, we could use the negative binomial or a continuous distribution such as the normal. Since the normal distribution theory is given in Hadley and Whitin (1963), the normal would seem to be the obvious choice. However, an iterative procedure is required, because the equation for the reorder point is a nonlinear function of the order quantity and vice-versa. Thus, it is necessary to make initial guesses for the Q and R, and then to refine them by substituting the latest estimate for Q into the equation for R, and then substituting this R into the equation for Q, etc. The procedure does converge rapidly in most cases, but is cumbersome.

Presutti and Trepp (1970) found that if demand is assumed to follow a Laplace distribution, the equation for Q is independent of R and no iteration is required. They show that the Laplace is similar to the normal, though it has a larger peak and longer tails for any specified standard deviation, a, as depicted in Figure 5-2. Again, our interest is not the difference in the

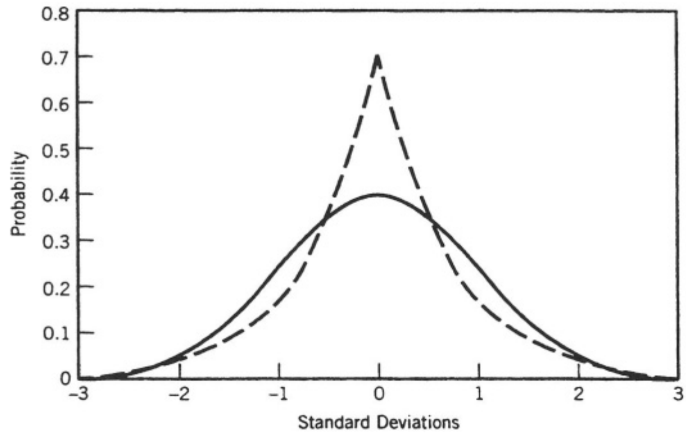


Figure 5-2. **Normal distribution (solid line) and Laplace distribution (dashed line) compared.**

individual probabilities, but the difference in the reorder point and order quantity from the two distributional assumptions. We will return to this consideration below.

We define the following variables for a particular item:

μ = expected demand during a lead (or resupply) time
T = constant lead time
σ = standard deviation of demand during a lead time
R = reorder point
k = protection level = $(R - \mu)/\sigma$
Q = order quantity

EBO(Q, k) = expected backorders at a random point in time, based on the two parameters Q and k instead of s as before. (Note the reorder point R can be computed from k).

The Laplace distribution for lead time demand is given by:

$$\text{lap}(x) = \left(\frac{\sqrt{2}}{2\sigma}\right)e^{-\frac{\sqrt{2}|x-\mu|}{\sigma}} \qquad -\infty < x < \infty \qquad (5.25)$$

Using the approach of Hadley and Whitin, the probability that the number of units backordered at time t is between y, $y + dy$ given that the inventory position (on-hand plus on-order minus backorders) is between L, $L + dL$, at time t $-T$ is:

$$\left(\frac{\sqrt{2}}{2\sigma}\right)e^{-\frac{\sqrt{2}|L+y-\mu|}{\sigma}}dy \tag{5.26}$$

This is because any stock on order at t - T must have arrived by time t, and any demands after time t - T cannot have been replenished. Hadley and Whitin show that the probability that the inventory position is between L, and L + dL is dL/Q. Therefore, Presutti and Trepp find that the probability that the number of units backordered is between y and y + dy is

$$\Pr\{y \le BO \le y+dy\} = \int_R^{R+Q}\left(\frac{\sqrt{2}}{2\sigma}\right)e^{-\frac{\sqrt{2}(L+y-\mu)}{\sigma}}dy\frac{dL}{Q} \tag{5.27}$$

Note that we have been able to delete the absolute values from Equation 5.25, because we consider values of y such that L + y is greater than μ. This is equivalent to requiring a positive safety level.

$$\Pr\{\text{BO} = y\} = \frac{1}{Q}\int_R^{R+Q}\left(\frac{\sqrt{2}}{2\sigma}\right)e^{-\frac{\sqrt{2}(L+y-\mu)}{\sigma}}dL \tag{5.28}$$

$$\Pr\{\text{BO} = y\} = \frac{.5}{Q}e^{-\frac{\sqrt{2}(y+k\mu)}{\sigma}}\left(1-e^{-\frac{\sqrt{2}Q}{\sigma}}\right) \tag{5.29}$$

and the expected backorders are:

$$\text{EBO}(Q,k) = \int_0^\infty \Pr\{\text{BO} = y\}dy = \frac{.25\sigma^2}{Q}e^{-\sqrt{2}k}\left(1-e^{-\frac{\sqrt{2}Q}{\sigma}}\right) \tag{5.30}$$

The annual costs of ordering and holding for an item at a site from Equation 4-57 in Hadley and Whitin [1963] are to be minimized with respect to Q and k:

$$\min\frac{\Omega m}{Q} + ic\left[\frac{Q+1}{2} + R - \mu + \text{EBO}(Q,k)\right] \tag{5.31}$$

subject to:

$$\mathrm{EBO}(Q,k) \le \beta$$

$$Q \ge 1$$

$$R \ge -1$$

where m = average annual demand on the next higher echelon support site

Ω = fixed order cost

ι = holding cost per unit per year

c = item cost

β = backorder target determined by VARI-METRIC (this must be the backorders for the portion of stock being resupplied from the next higher echelon)

Note that the first term of Equation 5.31 is the average cost per order times the average number of orders/year; the remaining terms are the average annual cost of holding a unit of the item times the average stock-on hand.[1] We substitute $k\sigma = (R - \mu)$ into Equation 5.31 so that the variables are now Q and k instead of Q and R. Then we want to minimize the sum of Equation 5.31 plus a Lagrange multiplier, λ, times the constraint Equation 5.30.

When this expression is differentiated with respect to k and set equal to zero, it can be solved for the Lagrange multiplier:

$$\lambda = \iota c\left(\frac{4Q}{\sqrt{2}\sigma\psi e^{-\sqrt{2}k}} - 1\right) \tag{5.32}$$

where we have defined

$$\psi = 1 - e^{-\frac{\sqrt{2}Q}{\sigma}} \tag{5.33}$$

When the same expression is differentiated with respect to Q and set equal to zero, it is possible to substitute for the Lagrange multiplier from Equation 5.32 and obtain:

[1] The average stock on hand can be computed from Equation 2.5, s = OH + DI - BO. In the place of the constant s, we substitute the term *inventory position,* IP, because it is not constant in the general case with $Q > 1$. Transposing terms and using expected values, we obtain E[OH] = E[IP] - E[DI] + E[BO].

$$0 = -\frac{\Omega m}{Q^2} + \frac{\iota c}{2} - \frac{\iota c}{\sqrt{2}\mu}\left(\frac{\psi}{Q} - \left(\frac{\sqrt{2}}{\sigma}\right)e^{-\frac{\sqrt{2}Q}{\sigma}}\right) \tag{5.34}$$

Note that Equation 5.34 is independent both of k (which defines the reorder point, R) and the backorder constraint, β. That is, of course, the reason that we are using the Laplace distribution. Equation 5.34 cannot be solved explicitly for Q. An iterative solution is possible, but an easier alternative is to recognize that in most cases ψ in Equation 5.33 is approximately one, so that Equation 5.34 can be written:

$$0 = -\frac{\Omega m}{Q^2} + \frac{\iota c}{2} - \frac{\iota c \sigma}{\sqrt{2}Q} \tag{5.35}$$

and

$$Q = \frac{\sigma}{\sqrt{2}} + \sqrt{\frac{2\Omega m}{\iota c} + \frac{\sigma^2}{2}} \tag{5.36}$$

Usually we constrain Q so that it does not exceed five years of demand, and it must be a positive integer. Then we can solve for k directly from Equation 5.30, because the backorders should equal the constraint β in order to minimize Equation 5.31:

$$k = -\frac{1}{\sqrt{2}}\log\frac{4Q\beta}{\sigma^2\left(1 - e^{-\frac{\sqrt{2}Q}{\sigma}}\right)} \tag{5.37}$$

where the constraint on R in the minimization of Equation 5.31 implies that $k \geq -(1 + \mu)/\sigma$, and $R = \mu + k\sigma$. At first glance it seems odd that the formula for k, the multiple of σ for safety level, should be preceded by a minus sign; but k is usually positive because we are taking the logarithm of a number which is typically less than one, and the result is a negative number.

Note that it is easy to solve these equations. We substitute into Equation 5.36 to obtain Q and use that value in Equation 5.37 to solve for k and the reorder point, R. Remember that the Q and R relate to the demand on the next higher echelon support site, excluding any demands that are repaired locally. By letting σ go to zero in Equation 5.36, we see that we have derived the Wilson lot size from Equation 1.1.

Suppose that the lead time is not constant. We used the constant lead time assumption in the derivation of expected backorders. If the lead time is allowed to vary, the estimate for expected backorders will not be precisely correct. Of course, it is an approximation anyway, because of our use of the Laplace distribution. But, if the lead time is not constant, then σ is the standard deviation of demand over the lead time including both demand variability over a fixed lead time and lead time variability. Thus, σ will be larger than in the constant lead time case (see Problems 4 and 5).

5.14 Numerical Example

Let's do a computation of the economic order quantity and reorder point at the depot for a specific item. Assume that demand is Poisson and:

Average annual depot demand = 20
Fraction not Repairable at Depot (NRTS) = 1
Unit cost = \$200
Average annual holding cost rate = 0.25
Order cost = \$300
Procurement lead time (constant) = 220 days

From the VARI-METRIC optimization that meets the target (availability, budget, or whatever), we use the expected backorders for this item at the depot of 0.1 as a target for our single-item computation.

We solve Equation 5.36 for Q first, and then substitute that value into Equation 5.37 for r. In order to compute Equation 5.36, we need the average annual demand on the support site (procurement source in this case). The annual demand on procurement, m, is just the annual depot demand multiplied by the NRTS fraction:

$$m = (20)(1) = 20$$

We need the standard deviation of lead time demand. Assuming Poisson demand and a constant lead time, the standard deviation of lead time demand is the square root of the lead time demand (see Section 1.10). The procurement lead time demand, μ, is just the average demand over the lead time:

$$\mu = (20)(220)/365 = 12.06$$

and the standard deviation of lead time demand is:

$$\sigma = \sqrt{\mu} = \sqrt{12.06} = 3.47$$

Finally, from Equation 5.36

$$Q = (.707)(3.47) + [(2)(300)(20)/(.25)(200) + (12.06)/2]^{0.5} = 18.13$$

which rounds to 18. Now we can substitute into Equation 5.37 for k:

$$k = -.707 \log \frac{(4)(18)(0.1)}{(12.06)[1 - e^{-7.33}]} = 0.396$$

and we find the reorder point, R, from the expression following Equation 5:37 as:

$$R = \mu + k\sigma = 12.06 + (0.396)(3.47) = 13.43$$

which rounds to 13.

Suppose we modify the problem so that the item is sometimes repairable, with a NRTS = 0.5; assume the average depot repair time is 30 days. Repeating the computation above, we find Q = 13 and R = 6.28 before rounding. However, the reorder point needs to be related to our inventory position, the sum of on hand plus due-ins minus backorders; we must add the depot repair pipeline which equals:

(Annual demand repairable at depot)(Depot Repair Time in Years)

$$= (10)(30)/365 = 0.82.$$

so that the reorder point is 7.10, or 7 after rounding.

There is one other little detail. The expected backorder target from VARI-METRIC was assumed to be 0.1 for the depot on this item. But in the case where some demand is repaired at the depot and some is reprocured, the 0.1 needs to be allocated between the two pipelines. We believe the most appropriate split is in proportion to the pipelines; thus, instead of 0.1 in the second case with some depot repair, the backorder target, used in the calculation of R, should be (0.1)(6.03)/(6.03 + 0.82) = 0.88. In this case, the impact is small, but when the NRTS fraction is fairly small, the expected backorder target for the procurement (or resupply) pipeline will be reduced significantly. On the other hand, if the NRTS fraction is small, the procurement demand will also be reduced, which reduces Q and R. For example, following the same calculations where the NRTS = 0.1 and the backorder target is reduced by the pipeline ratios to 0.045, yields $Q = 5$, and $R = 3$; the latter includes the depot repair pipeline of 1.48.

5.15 Item Criticality Differences

Our implicit assumption in all the modeling above has been that a backorder on any first-indenture item at a base is equally critical to the performance of the aircraft. As a first approximation, that appears to be reasonable in most cases. However, it is possible to refine that assumption. We have already discussed weighting the backorders differently by base when the number of aircraft vary by base, or when various parameters may vary by base (order and ship times, demand rates, repair times, etc.) and we want to meet specified base availability targets.

Suppose that there is an LRU that is only needed on some fraction f of the flights. Then it would be quite reasonable to assume that a backorder on that item should only have an imputed cost/day that is a fraction f as large. This is easily accomplished by allowing backorder weights not just by base, but by item. Of course, our interpretation of the term "availability" must be modified to take into account these partially available aircraft. While this seems like a reasonable procedure, it has not been adopted by any users to our knowledge.

Note that we do not need to worry about the criticality of lower indenture items or the criticality at support sites, because the backorders of LRUs at operating bases determine system availability completely. The impact of lower indenture items and other echelons on LRU base backorders is already taken into account.

Finally, we close this section on criticality with a true story. About forty years ago a contractor was building an inventory model for the Navy. The Navy promised to supply item criticality, but when the model was done the criticality had not been provided. The contractor was asked to use his best judgment. First, the contractor calculated the stockage policy on the assumption that criticality was proportional to unit cost, but was distressed to find that this led to a lot of stock on the high-cost items. Next the contractor assumed that criticality was independent of unit cost, only to discover that there was a lot of stockage on the low cost items. The final report embodied the contractor's best estimate that criticality was proportional to the fourth root of unit cost.

Of course, that is one of the "non-results" of inventory theory that has not survived the test of time. Obviously if we had a way of knowing how much to stock on the high cost items (and other parameters), we wouldn't need an inventory model in the first place.

5.16 Availability Degradation due to Maintenance

The availabilities in the models that have been developed increase with more spares, and could approach 100% with enough spares. However, even if we have infinite spare stock, there will be times when an aircraft is not operational because of the time to remove and replace a first-indenture item on the aircraft. In addition to maintenance time that is required to replace a failed item, there may be additional downtime due to periodic inspections.

It is easy to include this maintenance downtime to calculate Operational Availability as discussed in Section 2.14. The reason we have not done so previously is because the appropriate modification depends on the type of end item. For example, aircraft do not fly continuously. There is some turn-around time between flights for servicing the aircraft with fuel and other provisions. In reality the aircraft availability rates vary over time, decreasing during service periods. Normally we exclude downtime for servicing in the calculation of availability. Similarly if the aircraft fly in daylight hours only, we may not care about the availability rates at night.

In effect, the operational availability for aircraft depends on the flying program, because if there is sufficient time between flights or at night it may be possible to perform most of the remove-and-replace maintenance on first-indenture items without further degradation in availability, assuming that sufficient spares are available.

On the other hand if the end item is a system that operates 24 hours a day, there will be a further degradation in availability due to maintenance. If the system is down a fraction of time f for preventive maintenance, we must multiply the availability due to spares that we have calculated in Equation 5.24 by f. If we replace items as they fail, the overall availability is Equation 5.24 multiplied by f, the probability that no first-indenture item is down for maintenance, where

$$f = \prod_{i=1}^{I} (1 - \rho_i) \qquad (5.38)$$

and ρ_i is the probability that item i is undergoing remove and replace maintenance at a random point in time.

5.17 Availability Formula Underestimates for Aircraft

The availability formula of Equation 2.18 assumes continuous operation of the end-item. As noted in the previous section, we may not want to compute the maintenance availability to estimate operational availability when the end-items are aircraft with sufficient time to remove-and-replace items between sorties.

In addition, the availability formula of Equation 2.18 may understate aircraft availability when:

1) Sorties are longer.
2) Demand rates are high.
3) Aircraft fleets are large
4) There are long recovery periods between sorties.
5) There is some flexibility in scheduled takeoff times.
6) "Opportunistic" installation of spares is possible.

When sorties are longer and demand rates are high, it is more likely that there will be multiple failures. This does not usually result in an air abort, but it does prevent the aircraft from taking off until maintenance is performed. By "opportunistic" installation of spares we mean that maintenance is able to diagnose all failures, determine which items are needed, and only install the needed items on the aircraft if all items are available. There is no point in installing some of the items if that does not make the aircraft flyable, because those items may convert a down aircraft to flyable in the future (we are assuming that cannibalization is not normally performed). Of course, "opportunistic" replacement may not be feasible in all cases, because maintenance may find additional problems during "real-world" repair. Nevertheless, it does represent an upper bound on maintenance capability that is useful to examine.

Simulation is required to incorporate all these "real-world" sortie effects. Table 5-1 shows the percent reduction in the number of aircraft down from the number predicted by the availability formula of Equation 2.18. The data are invented so that only approximate conclusions can be made. The interested reader should write his own simulation (easy to do) using data from his application.

Table 5-1. % Reduction in Aircraft Down Estimated by the Availability Formula

Sortie Length	Random Replacement	Opportunistic Replacement
1-2 hours	0%	15.7%
4 hours	6.8%	23.1%
8 hours	14.6%	35.6%

5.18 Summary

In this chapter we have derived the VARI-METRIC theory for the combined multi-echelon, multi-indenture problem. Using this theory it is possible to calculate the cost-availability curve for aircraft at operating bases, and the associated stock levels for every item at every location corresponding to each point on the curve. The theory can also be used to calculate cost-availability curves for individual aircraft systems, such as propulsion. Cost-availability curves for different systems can be combined to give the same aircraft cost-availability curve, provided that there are no items that are common to different systems. When there are items common to more than one system, it is better to use VARI-METRIC on the combined systems.

Note that when VARI-METRIC is applied to the entire aircraft or multiple systems, it is not necessary to specify hierarchical information about the system on which each LRU is used. This is because we assume that any LRU backorder will decrease availability, regardless of the system on which it is used. In the next chapter we consider redundancy where the space station is available if K or more of the N identical copies of a system operate. We will find it necessary to compute the probability distribution of backorders for an item across the N copies of the system.

We do not consider redundancy in this chapter. However, the VARI-METRIC theory is easily adapted to cover redundancy in the simplest case when there is a single end item at each base (see Problem 6).

We discussed several generalizations to the VARI-METRIC assumptions. For example, a constant resupply time that does not vary by base was assumed in the derivation, and later relaxed. Similarly we have generalized the Poisson assumption to the case where demand rates vary over time and to the case where failures are due primarily to wear out.

We have shown how it is possible to modify the theory for SRUs that are common to more than one LRU parent. Using backorder targets by item and location from VARI-METRIC, we have developed a procedure for computing the reorder point and order quantity. There was a discussion of techniques for modeling criticalities that vary by item, and for including in availability the degradation due to maintenance remove-and-replace time.

5.19 Problems

1. Integrate Equation 5.28 and obtain Equation 5.29. Also verify the integration in Equation 5.30.
2. Verify Equations 5.32 and 5.33.

3. Calculate the mean and variance for $E(X_{00})$ and $\text{Var}(X_{00})$ from Equations 5.16 and 5.17 when there are two SRUs where each $f_{i0} = .5$ and $m_{i0}T_{i0} = 1$ for the LRU and each SRU. The SRU stock levels at the depot, s_{i0}, are each one. Use Table 4.2 to show that when demand is Poisson $E(X_{00}) = 1.3679$ and 1.4324 and that when demand has a variance-to-mean, V, of 3 over the pipeline $E(X_{00}) = 1.5774$ and $\text{Var}(X_{00}) = 4.3335$. Use Table 4.5 to show that when demand has a binomial distribution with $V = .5$ over the pipeline $E(X_{00}) = 1.2500$ and $\text{Var}(X_{00}) = .7188$. What do you conclude about the sensitivity of these pipeline means and variances with respect to the differences in the demand processes?

4. In Section 5.13 the lead time was assumed to be constant. When the lead time has a standard deviation σ_L, show that the variance in lead time demand due to both demand variability and lead time variability is:

$$\sigma^2 = (m\sigma_L)^2 + V\mu$$

where m is the average annual demand, V is the variance-to-mean ratio of demand, and μ is the expected lead time demand. What does this become when V is given by a power curve relationship such as Equation 4.22 for some parameters a and b: $V = 1 + am^b$? *Hint*: Use Equation 5.2 where the conditional variable is the lead time, T, and x is demand over the lead time

$$\text{Var}(X) = \text{Var}[E(X|T)] + E\,[\text{Var}(X|T)]$$

5. Effect of variability on the economic order quantity, Q.

a) Assume that demand is Poisson and $m = 10$ demands/year, NRTS = 1, procurement lead time = 0.5 years, $V = 1$, ιc (holding cost) = \$100/year, Ω (fixed order cost) = \$100 in Equation 5.36 and assume that the lead time is constant. Show that before rounding to an integer the value of $Q = 6.32$.

b) Assume that the standard deviation of the lead time, σ_L is 0.2 years, and use the result of Problem 5.4 above to show that the standard deviation in lead time demand, σ, equals 3. Recompute Q from Equation 5.36 and show that it now equals 7.07 before rounding.

c) Assume the conditions of part a and b except that the variance-to-mean ratio of demand, V, is given by the power relationship of Equation 4.22. Show that the standard deviation in lead time demand, σ, now equals 3.33 and Q from Equation 5.36 now equals 7.39 before rounding.

Note that changes in the standard deviation of lead time demand result in less than proportional changes in the order quantity Q. Thus Q is more robust than σ to data errors.

6. Consider a nuclear power plant application where the single end item is the power plant itself. The power plant must shut down (or operate at reduced power) unless at least K of N water circulating pumps are operating.

Show that with a simple change in Equation 5.24, it is possible to include redundancy (no cannibalization). (Assume that the demand rate for an item is independent of K, otherwise known as the cold-standby case. See problem 6 of Chapter 8 for the cannibalization case.)

7. (Research) Attempt to develop an analytic model and verify with simulation the behavior of the variance in the pipeline distribution for a single item in the multi-echelon case when the order and ship times are allowed to vary by base. Consider rules that may be more effective than first-come, first-serve for allocating depot shipments to bases.

8. Use VARI-METRIC theory to recalculate the example in Section 3.4. In particular show that the row for a depot stock level of 2 in Table 3.3 becomes 1.9240, 1.6114, 1.2988, 0.9862, 0.6736, 0.3610, 0.2995

Chapter 6

MULTI-ECHELON, MULTI-INDENTURE MODELS WITH PERIODIC SUPPLY AND REDUNDANCY

In the United States there is more space where nobody is than where anybody is. That is what makes America what it is.

– Gertrude Stein

6.1 Space Station Description

The theory in this chapter was motivated by a need to develop a stockage model for Space Station *Freedom,* the large multi-national effort being managed by NASA that was launched in the mid-nineties. The most important difference between this application and those considered in the previous chapters is that here the operating base cannot place a resupply order at any time. Resupply is periodic, depending on the frequency of the space shuttle flights. There will be several a year, but depending on the item and whether it needs pressurized or unpressurized storage, it may only be eligible for certain resupply flights. This periodic supply modeling is applicable to other problems such as a base in Antarctica which can only be resupplied during the summer months or ships which can be resupplied only when they return to port.

A Space Station is comprised of many different types of systems such as Electric Power, Laboratory Module, Habitation Module, etc. A given system type will usually have significant redundancy in its design. After all we are talking about systems that may be critical for life support, and which are designed to operate for thirty years.

Instead of a fleet of aircraft, we are concerned here about systems. For a particular system type with N identical copies, we assume that at least K units must be operating for the Space Station to operate. For each first-indenture, orbital replacement unit (ORU) in a system we assume that at least z of the Z identical copies must operate for the system to operate. Thus, we are modeling redundancy at two levels. This enables us to consider a wide variety of design configurations and maintenance policies ranging from no cannibalization to full cannibalization.

Cost is not the only resource that we want to consider in the optimization. There are weight and volume constraints in space that must be considered when we determine what to buy and whether to keep it in space or on the ground.

Finally, most ORU failures are assumed to be due to random causes and they are usually repairable on the ground. The theory below is based on these assumptions initially. Then it is generalized in Chapter 7 to cover those ORUs whose failures are due primarily to wear out, e.g. batteries. Also, some percentage of failures on an ORU, whether due to random causes or wear out, can not be repaired. There must be additional spares to cover these condemnations.

The theory developed in this chapter is for a specific application, and this provides important motivation. However, it is important for the reader to bear in mind that the techniques for modeling periodic resupply and redundancy are very powerful and can be used in other problems.

6.2 Chapter Overview

In this chapter we describe the "physics" of the Space Station resupply and repair processes. Because resupply is periodic, the availability on the station is highest after a resupply and declines over the cycle. Similar to previous chapters, on each item we find the probability distributions for the number of units in repair on the ground and the number of units broken on orbit. Subtracting this from the total number of spares, we determine what is available to resupply the station via the space shuttle, and this allows us to calculate station availability at any time t during the next cycle.

We begin with a single ORU and compute the probability distribution for the number of systems down due to that ORU. Then we show how to combine ORUs to calculate the probability distribution for the number of systems down due to all ORUs comprising a system. Because of redundancy within a system, a "hole" for an ORU does not always cause a system to be down. At the system level we assume there are N identical copies of the system of which only K must be operating at any point in time. Thus, we have to compute the availability taking this second type of redundancy into account as well.

The traditional approach used by NASA to compute stock levels is described. The defects of this approach are examined, and the dramatic improvements obtainable with the optimal theory are illustrated with several examples. Reliability block diagrams are discussed, and converted to the appropriate parameters for the theory. We show that it is possible to evaluate the availability of different reliability configurations. This should be useful to system designers even before considering optimal stockage policies.

We explain how marginal analysis can be used in this problem to obtain optimal stockage policies for what to stock on the ground and on orbit. The impact of cannibalization is assessed. We describe the state-of-the-art in redundancy modeling and present some important extensions to the theory. The chapter concludes with a discussion of how the theory can be applied. Some important topics are left as problems at the end of the chapter: (1) why earlier approaches to modeling redundancy understate availability, particularly when the number of systems is small; (2) how to extend the model to also optimize the multi-indenture stockage of SRUs on the ground.

A number of generalizations of the theory are deferred to the next chapter to avoid distracting the reader. More examples of how the theory can be applied are found there as well.

6.3 Maintenance Concept

The maintenance concept for a space station is that when an ORU fails on the station, it is removed and replaced by a spare if one is available on the station. The next space shuttle will bring some resupply of ORUs, and take the failed ORUs back to the ground for repair. Even though resupply is periodic, we assume that information on orbital failures is sent to the ground continuously.

Most ORUs are repairable on the ground. These are taken into a shop or sent to a contractor where the failed SRUs are replaced on the ORU. The time to repair a specific ORU is assumed to have a probability distribution, depending on the complexity of the repair and the availability of SRUs.

Our model needs to be multi-echelon for the ORUs, specifying the appropriate stock level in space and on the ground for each item. It needs to be multi-indenture as well, where the optimal mix of SRU ground stock is determined.

6.4 Availability as a Function of Time during the Cycle

The measure of effectiveness is the availability, the probability that the Space Station (or some system type or group of system types)[1] is not down due to spares multiplied by 100 and expressed as a percent. As contrasted with our previous models, however, the availability here does not stay constant. After a resupply shuttle mission, the availability is highest and then it decreases until the next shuttle resupply flight as depicted in Figure 6-1. It decreases because failures occur during the cycle, and these are removed and replaced with spares. The probability that we will run out of some type spare, and fall below the minimum of K operating systems of some type is greatest just before the next resupply.

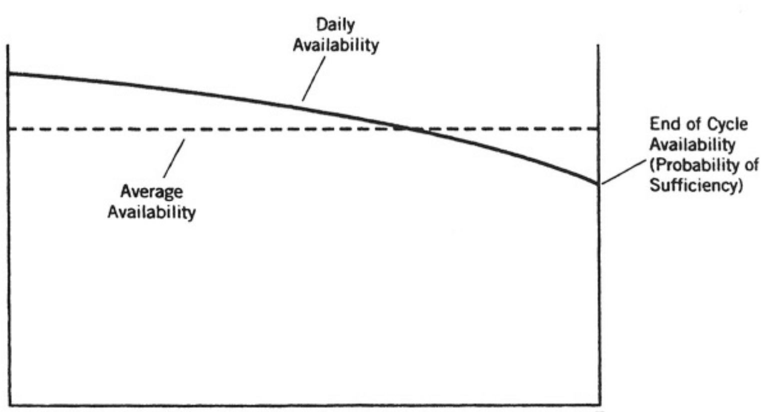


Figure 6-1. **Availability on the space station: different measures.**

Since availability is not a constant, we have to decide what point in time we are interested in. For some systems, particularly those related to safety of flight, it seems logical to be most concerned about those times at which availability reaches its lowest value - at the end of each cycle just before the next shuttle arrives.

[1] For a system with N copies, there must be some number $K \leq N$ systems operating. K and N will vary by system type, of course.

For other systems, such as the Electric Power System, we may be more interested in average availabilities. This is because it is not critical that there be 100% power throughout each cycle. In the Electric Power Generation System there were planned to be eight different "strings" ($N = 8$), each of which would produce one-eighth of the total power. Thus, we may want a power generation profile that estimates the average fraction of each cycle at 100% power (8 of 8 strings operating), at 87.5% power (7 of 8 strings), at 75% power (6 of 8 strings) etc. This implies that we need the capability to assess availability at any time point in the cycle.

6.5 Probability Distribution of Backorders for an ORU

We begin with a single ORU, assumed to be repairable always, and demand on the station during a cycle of some fixed length, T years. Random failures will be assumed throughout this chapter, so that demand is from a Poisson process with a constant average annual mean, m:

$p(x|mT)$ = probability that demand over the cycle is x where $x = 0, 1, 2,...$ when the mean demand is mT

s = spare stock for the ORU composed of the orbital stock level s_O and the ground stock level s_G. This means that we send up enough stock on each shuttle to restore the orbital spare stock to s_O, if possible. Of course, there may be occasions when we are not able to restore the orbital spare stock to s_O due to the number of failed units of the ORU.

Our objective is to compute the probability distribution for the number of backorders or "holes" for the ORU at an arbitrary time t during a cycle where $0 \leq t \leq T$. We need to compute not only the probability of sufficiency (or the expected backorders as in previous chapters), but the probability distribution for the number of backorders, because of our modeling of redundancy, to come later.

Consider the time 0 at which the shuttle is to begin its resupply mission, as shown in Figure 6-2. If mT is the mean demand during a cycle and x denotes the number of demands during the cycle, then $p(x|mT)$ is the probability distribution for the number of ORUs that have failed in space during the cycle ending at time 0 and which will be taken back to ground on the shuttle return trip (we assume that there is ample space on the shuttle for all failed ORUs).

Suppose that some failures cannot be repaired during the time between a shuttle return to ground and the next shuttle flight. Let m_2T be the mean for the ORU demand that requires at least two cycles to repair (cannot return on the next shuttle), m_3T the fraction of the mean that requires at least three cycles to repair, etc. up to some maximum number of cycles, *n*. Logically these *m*'s must form a nonincreasing sequence. Each demand from the original Poisson process with mean m has a probability distribution for the number of periods that will be required for the repair. This divides the original Poisson process into n independent Poisson processes, one for each repair length.

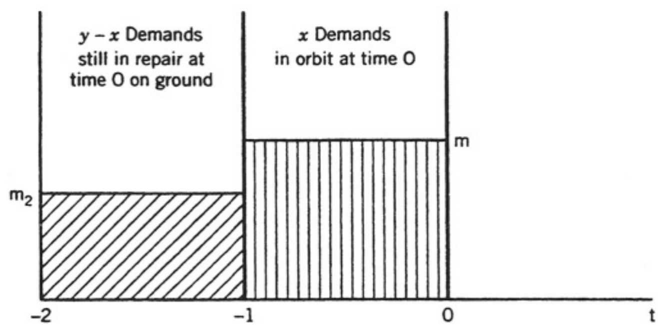


Figure 6-2. **Combinations of demand that result in *y* broken units at time 0 (when shuttle begins resupply mission).**

Figure 6-2 depicts the case where m_3T, the mean demand requiring three periods or more to repair, is zero. As in our previous chapters on repairable items, our task is to compute the probability distribution for the number of units in repair or not available for issue, but now it is at the time 0, when the shuttle departs for the station. Since the total spare stock *s* is constant, a knowledge of the number of units due in enables us to compute the number of units that are available to meet demand. Define

$\Pr\{DI = y\}$ = probability that the number of units due in (in repair on the ground or failed in orbit) at time 0 is y where $y = 0, 1, 2, \ldots$

As seen in Figure 6-2, for every value of *y* of interest we need to compute the probability of *x* demands during the orbital cycle ending at time 0 and *y* - *x* demands during the previous cycle. The latter group of *y* - *x* broken ORUs was returned to ground on the previous shuttle, but could not be repaired by time 0. This sum of probabilities is called a *convolution*:

$$\Pr\{DI = y\} = \sum_{x=0}^{y} p(x|mT)\,p(y-x|m_2T) \qquad y = 0, 1, 2, \ldots \qquad (6.1)$$

Note that we assume, as in the models of the previous chapters, that there is no queueing for repair on the ground; there is ample repair capacity. For m_3 not equal to zero, Equation 6.1 is applied again where the $p(x|mT)$ terms are replaced by the Pr{DI = x} terms just computed and the $p(y - x|m_2T)$ terms are replaced by $p(y - x|m_3T)$, etc.

In fact it is not necessary to calculate the convolution in Equation 6.1 or its generalization when $m_3 > 0$, $m_4 > 0$ etc. We know from Problem 12 of Chapter 4 that Pr{DI = y} is Poisson with mean $(m + m_2 + m_3 \ldots)T$, and thus can be calculated directly. Later we will generalize the Poisson demand assumption on the right-hand side of Equation 6.1 to the negative binomial and binomial distributions (for variance-to-mean ratios greater than one and less than one, respectively). From Problems 13 and 14 of Chapter 4 we know that both of these probability distributions are convolutions also, and thus Pr{DI = y} can be computed directly, provided that the variance-to-mean ratio is constant.

Now we are ready to compute the probability that there is exactly one backorder Pr{BO = 1} for the ORU at time t, $0 \le t \le T$, during the next cycle. There will be exactly one backorder if we had s_G or less in repair on the ground and failed in orbit at the time the shuttle left for the Space Station at time 0 and $s_O + 1$ demands during the orbital cycle up to time t. This is because with s_G or less broken, we were able to fill the spare stock on orbit to the full stock level s_O with the shuttle flight that arrived at the beginning of the cycle (the shuttle arrives at some time later than 0, but the time interval is short and can be ignored for most purposes). Incidentally, the number of spares sent up in the shuttle is not s_O, but x where x is the number that are broken in orbit; there will be s_O good spares in orbit after the shuttle arrives.

There will be exactly one backorder also if we had $s_G + 1$ broken at time 0 and s_O demands during the orbital cycle up to time t. In this case we were able to fill the spare stock on orbit only to the level $s_O - 1$ with the shuttle flight. The other combinations yield a sum of terms:

$$\Pr\{BO = 1\} = p(s_O + 1|mt)\Pr\{DI \le s_G\} + p(s_O|mt)\Pr\{DI = s_G + 1\}$$

$$+ p(s_O - 1|mt)\Pr\{DI = s_G + 2\} + \ldots + p(0|mt)\Pr\{DI = s_O + s_G + 1) \tag{6.2}$$

and

$$\Pr\{BO = x\} = p(s_O + x|mt)\Pr\{DI \le s_G\} + p(s_O + x - 1|mt)\Pr\{DI = s_G + 1\}$$

$$+ p(s_O + x - 2|mt)\Pr\{DI = s_G + 2\} + \ldots + p(0|mt)\Pr\{DI = s_O + s_G + x)$$

$$x = 1, 2 \ldots \tag{6.3}$$

Note that the mean demand through time t during a cycle, where $0 \le t \le T$, is just mt. The first term on the right-hand-side of Equations 6.2 and 6.3 differs from the subsequent terms, because it contains a cumulative probability for the number of units due in. Even though the probabilities for the number of units due in from Equation 6.1 are Poisson probabilities. Equations 6.2 and 6.3 are not convolutions that can be evaluated immediately as in Equation 6.1. This is because the arguments for $p(\ )$ and $\Pr\{\ \}$ have different ranges and there are some cumulative probabilities.

We need the expression for $\Pr\{\mathrm{BO} = 0\}$ as well. There will be no backorders if we had s_G or fewer broken at time 0 and s_O or fewer demands during the orbital cycle up to time t. Also if we had $s_G + 1$ broken and $s_O - 1$ or fewer demands during the orbital cycle up to time t, etc. The complete expression is:

$$\Pr\{\mathrm{BO} = 0\} = P(s_O | mt)\Pr\{\mathrm{DI} \le s_G\} + P(s_O - 1 | mt)\Pr\{\mathrm{DI} = s_G + 1\}$$

$$+ P(s_O - 2 | mt)\Pr\{\mathrm{DI} = s_G + 2\} + \ldots + P(0 | mt)\Pr\{\mathrm{DI} = s_O + s_G) \qquad (6.4)$$

Note that Equation 6.4 differs from Equation 6.2 because it contains cumulative Poisson probabilities, $P(\ )$, rather than Poisson probabilities $p(\ )$ for the number of demands.

6.6 Probability Distribution for Number of Systems Down for an ORU

Now we want to compute the probability distribution for the number of units of a system that are down for this ORU. Suppose that the quantity of this ORU on a system is Z, and that z of the Z must operate for the system to operate. Obviously z must be less than or equal to Z. When z can be less than Z, there is redundancy within a system.

We want to consider redundancy at the system level as well by specifying that at least K units of the total N copies of the system must operate. Our objective is to compute the probability distribution for K units of the system up (i.e functional) when one ORU is considered. Then we generalize this to all ORUs comprising the system. To do this we must consider a specific configuration where for convenience we assume that the first K systems must be up and the last N - K are down (i.e. not working). Later we consider the number of different ways that K systems can be selected from N.

The hypergeometric distribution[1] gives the probability distribution that N systems have $x_1, x_2, \ldots x_N$ backorders respectively, given a total of

$$y \equiv \sum_{j=1}^{N} x_j$$

backorders across all N systems:

$$\text{hyp}(x_1, x_2, \ldots, x_N \mid y) = \frac{\binom{Z}{x_1}\binom{Z}{x_2}\cdots\binom{Z}{x_N}}{\binom{NZ}{y}} \tag{6.5}$$

This is because there are Z locations in each system, and x_1 of these are to have backorders on system 1, x_2 on system 2, etc. This is conditional on the fact that there must be y backorders among all NZ possible locations.

The probability distribution for the number of backorders, Pr{BO = y}, was computed above in Equation 6.3. The probability that systems 1, 2,. . .i are up and systems $i + 1$, $i + 2$, . . . N are down can be written:

$$S(i) = \Pr\{\text{BO} = y\} \sum_{W} \text{hyp}(x_1, x_2, \ldots, x_N \mid y) \qquad K \le i \le N \tag{6.6}$$

where W is the set of hypergeometric probabilities such that the backorders on system j, x_j, have the properties that:

$$0 \le x_j \le Z - z \ \text{ for } j \le i; \qquad Z - z \le x_j \le Z \ \text{ for } j > i.$$

In other words for each value of y, we must add all those hypergeometric probabilities with the property that the first i systems are up (because no more than the allowable $Z - z$ units of the ORU are down) and the last $N - i$ systems are down.

Equation 6.6 looks very forbidding, but it is actually quite straightforward. We need to compute the probability sums over W for specific values of N, N - 1, . . . K as a function of y on an ORU. There is a

[1] It is not obvious that the hypergeometric is a valid probability distribution whose terms sum to one. The reader was asked to prove this in Problem 13 of Chapter 4.

lower bound on y that depends on z and Z for the ORU, below which the sum of the hypergeometric probabilities comprising the set W is 1 for $i = N$ and zero for all i, $K \le i \le N$ (see Problem 6).

For values of y between the bounds we must compute the hypergeometic probabilities for N, $N - 1$, $N - 2, \ldots K$. But, these probabilities are independent of the demand rates and stockage policy. Since they depend on z and Z only, two ORUs with the same z and Z have identical hypergeometric probabilities. Thus, the probability sums over W for specific values of N, $N - 1$, $N - 2, \ldots K$ as a function of y between the bounds can be computed once and used over and over with values of $\Pr\{\mathrm{BO} = y\}$ that differ by ORU and change as the stock level is augmented.

As an example consider one ORU where $z = 1, Z = 2$, and $N = 3$. We will work out the case where the number of systems operating, K, equals 2 which means that systems one and two are up and system three is down. The set W is empty for zero or one backorder, because there must be at least two backorders for a system to be down.

With two backorders the only hypergeometric probability in the set W is:

$$\text{hyp}(0,0,2) = \binom{2}{0}\binom{2}{0}\binom{2}{2}\Big/\binom{6}{2} = 1/15$$

With three backorders, there are two hypergeometric probabilities in the set W:

$$\text{hyp}(0,0,2) + \text{hyp}(1,0,2) = \binom{2}{0}\binom{2}{1}\binom{2}{2}\Big/\binom{6}{3} + \binom{2}{1}\binom{2}{0}\binom{2}{2}\Big/\binom{6}{3} = 1/5$$

With four backorders, there is only one hypergeometric probability in the set W:

$$\text{hyp}(1,1,2) = \binom{2}{1}\binom{2}{1}\binom{2}{2}\Big/\binom{6}{4} = 4/15$$

When there are more than four backorders, it is not possible to have systems one and two up and system three down. Applying the same approach for the other values of K yields Table 6-1.

Table 6-1. **Hypergeometric Example" \f a**

Number of Back Orders, y	Probability That a Specific Configuration of Exactly i Systems of N Will Operate $i = 3$	$i = 2$	$i = 1$	$i = 0$
0	1	0	0	0
1	1	0	0	0
2	4/5	1/15	0	0
3	2/5	1/5	0	0
4	0	4/15	1/15	0
5	0	0	1/3	0
6	0	0	0	1

To calculate the probabilities that the first i systems are up, $S(i)$, we multiply the hypergeometric probabilities in the corresponding column of Table 6-1 by the backorder probabilities in accordance with Equation 6.6. Thus,

$$S(2) = \Pr\{BO = 2\}/15 + \Pr\{BO = 3\}/5 + 4\Pr\{BO = 4\}/15.$$

6.7 Probability Distribution for Number of Systems Down

We begin by showing how two ORUs are combined; the procedure is repeated for as many ORUs as required. The problem in combining ORUs is that for a given system to be up, it must be up for each ORU. We will use subscripts on $S(.)$ to designate the particular ORU, and suppress the subscript to indicate a group of ORUs. The first three equations are:

$S(\mathrm{N})$ = probability that all N systems up

$$= S_1(N)S_2(N)$$

$S(N - 1)$ = probability that systems 1, 2, . . . N - 1 are up and N is down

$$= S_1(N - 1)S_2(N) + S_I(N)S_2(\mathrm{N} - 1) + S_1(N - 1)S_2(N - 1)$$

$$= S_1(N - 1)[(S_2(N - 1) + S_2(N)] + S_1(N)\mathrm{S}_2(N\text{-}1)$$

$S(N - 2)$ = probability that systems 1, 2, . . . N - 2 are up and N - 1, N are down)

$$= S_1(N - 2)[(S_2(N - 2) + 2S_2(N - 1) + S_2(N)]$$

$$+ 2S_1(N-1)[(S_2(N-2) + S_2(N-1)) + S_1(N)S_2(N-2) \qquad (6.7)$$

Note that there is a triangular symmetry here in the equations, similar to that in the development of the backorder probabilities in Equations 6.2 and 6.3. The last equation in the set above for $S(N-2)$ is more complicated because of the multiplicative factors of 2. The first factor of 2 arises from the fact that if two systems are down because of ORU 1 and one of these same systems is down for ORU 2, there are two possible ways of selecting the latter system; the second factor of 2 is a mirror image case. The third factor of 2 arises from the fact that $S_1(N-1)$ is the probability of having a specific system down due to ORU 1 and $S_2(N-1)$ is the probability of having a (different) specific system down due to ORU 2, but there are two ways of selecting these. The equations become more complicated as the number of systems allowed to be down in Equation 6.7 increases. Of course, N is not very large for most systems and the number K that must operate is positive, so that $N-K$, the number allowed to be down, will be quite small. Furthermore, we need never compute the probabilities for less than K systems. (see Problem 8)

These calculations for combining multiple ORUs are not required for the example of Table 6-1, because there was only one ORU.

6.8 Availability

The procedure above is repeated to include all ORUs in a system. To calculate the probability that any group K of N systems are operating, we must take the probability of a specific configuration such as the first i operating as defined by $S(i)$ and multiply by the number of different ways that i good systems can be drawn from a total of N. Thus, the availability, A, is 100 times the probability that any i or more of N systems are operating:

$$A(i) = 100[S(N) + \binom{N}{N-1} S(N-1) + \ldots + \binom{N}{i} S(i)] \qquad K \le i \le N \qquad (6.8)$$

Consider the example of Table 6-1 again. The probabilities in each row of the table must add to one when each entry is multiplied by the appropriate combinatorial factor for the column. For example, in the column headed $i = 3$ the combinatorial factor is one, because there is only one way of picking 3 good systems from 3 systems; in the column headed $i = 2$ it is three, because there are 3 ways of picking 2 good systems from 3. Thus, in the row for two backorders $4/5 + 3(1/15) = 1$. This can be used as a check on the probabilities in Table 6-1 and in the computation of availabilities from Equation 6.8. For the single ORU considered in the example of Table 6-1,

$$A(2) = 100[S(3) + 3S(2)] = 100[\Pr\{BO = 0\} + \Pr\{BO = 1\} + \Pr\{BO = 2\}$$

$$+ \Pr\{BO = 3\} + 4 \Pr\{BO = 5\} /5]$$

6.9 Numerical Example for one ORU

Let's consider the ORU with the redundancy configuration from Section 6.6. We want to calculate the backorder distribution where we will assume that the average demand over a resupply cycle, $mT = 1$. The demand which takes one or more cycles to repair plus the demand which takes two or more, etc., $(m + m_2 + m_3 \ldots)$ multiplied by T will be assumed to equal 2. From the discussion following Equation 6.1 (and proved in Problem 12), we know that when demand is Poisson Equation 6.1 is just a convolution which becomes:

$$\Pr\{DI = y\} = p(y|(m + m_2 + m_3 \ldots)T) = p(y|2) \qquad y = 0, 1, 2, \ldots$$

We use Equation 6.3 to calculate the probability distribution for backorders greater than 0 and Equation 6.4 for zero backorders. The time of interest, t, will be taken as the end of the resupply cycle, T, just before the next shuttle arrives. We will assume that the orbital spare stock, $s_0 = 2$, and the ground spare stock, $s_G = 0$. Thus, Equation 6.3 for $x = 1$ becomes:

$$\Pr\{BO = 1\} = p(3|1)p(0|2) + p(2|1)(p(1|2) + p(1|1)p(2|2) + p(0|1)p(3|2)$$

$$= p(3|3) = 0.2240$$

where the final simplification comes about because when $s_G = 0$ we have another convolution.

Similarly, we find:

$$\Pr\{BO = 2\} = p(4|3) = 0.1680 \qquad \Pr\{BO = 5\} = p(7|3) = 0.0216$$
$$\Pr\{BO = 3\} = p(5|3) = 0.1008 \qquad \Pr\{BO = 6\} = p(8|3) = 0.0081$$
$$\Pr\{BO = 4\} = p(6|3) = 0.0504 \qquad \Pr\{BO = 0\} = 0.4232$$

where Equation 6.4, which is not a convolution, is used for the last quantity above, $\Pr\{BO = 0\}$.

The probability distribution for the number of operating systems can be seen from Equation 6.8 as:

Pr{all 3 operate} = $S(3) = 0.8219$

Pr{any 2 operate} = 3S(2) = 0.1344
Pr{any 1 operates} = 3S(1) = 0.0316
Pr{none operate} = S(0) = 0.0081

and the expected number of systems operating is 2.78. The expression for S(2) was given at the end of Section 6.6. The reader is asked to derive the expressions for S(3), S(1), and S(0) from Table 6.1 as Problem 6.4.

Isaacson et. al. (1988) have modelled redundancy in a group of aircraft by calculating the probability that a system picked at random is not down. Then $A(i)$, the availability of i or more systems, is estimated by using a binomial distribution for the probability of i or more systems being up. From Table 6-1 the probability that a system picked at random is not down equals:

$$\Pr\{BO=0\} + \Pr\{BO=1\} + 14\Pr\{BO=2\}/15$$

$$+ 4\Pr\{BO=3\}/5 + 9\Pr\{BO=4\}/15 + \Pr\{BO=5\}/3$$

Denoting this value by ρ, the binomial probability of all 3 systems operating is ρ^3, any 2 systems operating is $3\rho^2(1-\rho)$, any 1 system operating is $3\rho(1-\rho)^2$, and none operating is $(1-\rho)^3$. These values are 0.7925, 0.1917, 0.0154, and 0.0004 respectively. The expected number of systems operating is still 2.78 as in the hypergeometric theory, but note how the binomial assumption reduces the variability dramatically. If we are really concerned about the extreme cases such as no systems operating, the binomial theory estimates a probability of only 0.0004 whereas the more exact theory estimates 0.0081 or twenty times as large.

The binomial approximation is poor because of the assumption of an infinite population of systems. It assumes sampling with replacement.

6.10 Optimization

We have shown how to evaluate any stock levels s_O, s_G for each ORU in a group that comprise a system with redundancy at two levels. In order to optimize, we need only be able to evaluate any policy, and develop an efficient procedure to generate solutions. As in the steady-state models developed in earlier chapters, marginal analysis techniques are utilized here.

Since the availability of several different system types is the product of the availabilities, we compare the increase in the logarithm of availability divided by the increase in cost for every possible augmentation (i.e. an increase in s_O or an increase in s_G for every item in a given system). After we have generated all of the potential solutions for one system type, we discard any that are not on the convex hull. Finally, we merge the solutions for different system types to generate the cost-availability curve.

6.11 Multiple Resource Constraints

As noted earlier, cost is not the only resource of interest here. Orbital weight and volume are considerations as well. This means that when we look at the ground resources used in augmenting s_G we divide by the item cost, c; when we look at the orbital resources used in augmenting s_O we divide by $c + \lambda_1 w + \lambda_2 v$ where the λ 's are Lagrange multipliers and *w* and *v* are the ORU weight and volume.

By varying the Lagrange multipliers we obtain a set of solutions. Suppose that we want 95% availability and the total volume from a particular set of λ's exceeds the allowable volume for storage. Then we must increase λ_2, the price of volume. Similar considerations apply to the weight constraint. Of course, the set of alternative solutions is finite. It may not be possible to achieve a 95% availability with a given volume and weight constraint. In such a case we must consider redesign, a change in our constraints, or a relaxation of our availability target.

Figure 6-3 provides actual system output from the photo-voltaic (PV) module, whose detail data are given in Table 6-2. Since there was no volume data by item at this time, we show the 90% and 95% availability tradeoff curves for investment cost in spares (on orbit plus on the ground) versus on orbit spares weight. It is quite typical to find that part of each curve is nearly vertical and part is nearly horizontal (a quick look at Table 6-2 shows a low correlation between cost and weight). For example, if our objective had been 95% availability and weight were the only resource considered, our optimal solution would have been the rightmost point on the graph - a weight of 16.9 thousand pounds. The associated spares cost of this policy is \$168.7 million. Figure 6-3 shows that if we relax the weight constraint slightly to 17.2 thousand pounds (a 1.8% increase in weight), there is a large decrease in cost to \$146.2 million (a 13.3% decrease in cost). Even when management is primarily concerned with one resource, these tradeoff curves provide important insight.

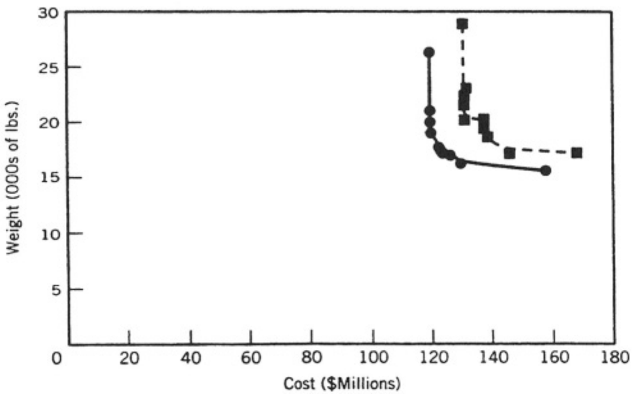


Figure 6-3. Constant availability curves. Boxes (□) signify 95% curve; Circles (○) signify 90% curve.

Table 6-2. PV Module Input Data

		System Name		K	N		
		PV Module		8	8		
ORU Name	**Cost ($000)**	**Wt**	**Vol**	**z**	**Z**	**IS**	**Annual Demand per Unit**
1	4047	50	0	1	1	1	.00832
2	93.6	64	0	1	1	1	.01180
3	130	85	0	1	1	2	.02666
4	2791	1316	0	1	1	3	.10344
5	240	334	0	1	1	4	.19356
6	240	299	0	3	3	8	.17648
7	555	158	0	1	1	4	.22220
8	15.6	93	0	1	1	4	.17144
9	1979	168	0	3	3	7	.16216
10	494	182	0	1	1	3	.11216
11	308	40	0	1	1	3	.08220
12	326	84	0	1	1	4	.23528
13	149	23	0	1	1	4	.21052
14	187	72	0	1	1	1	.03360
15	133	20	0	1	1	2	.03332
16	1604	30	0	1	1	2	.03332
17	7.8	12	0	1	1	3	.13188
18	19	231	0	1	1	1	.02560
19	6407	491	0	1	1	2	.09916
20	2000	38	0	1	1	3	.16664
21	664	147	0	1	1	3	.16436
22	6	2	0	4	4	4	.01320
23	630	167	0	1	1	1	.09375

6.12 REDUNDANCY BLOCK DIAGRAMS

Figure 6-4 shows a redundancy block diagram for the communication and transmission system in the space station. Numbers have been substituted for the names of each of the eight different orbital replacement units. The purpose of this section is to show how to convert a redundancy block diagram into several different system types where the modeling of a particular system type was described above. For a particular type of system there are N identical units of which K must operate, and in each of the K operating systems at least z of the Z ORU's of each kind must operate, (z and Z can vary by ORU within a system and K and N can vary from one system to another).

From a modeling perspective the overall problem can be thought of as four system types in series, and the results written down immediately as Table 6-3. Each system type in Figure 6-4 is enclosed in a dotted box for clarity.

Table 6-3 is not the only way to represent the physical system. Since system 2 contains only one item, the data for ORU 5 could be entered as $K = N = 1$, $z = 1$, $Z = 2$ and the same result obtained; similarly for system 3. Or, systems 2 and 3 could be combined into one system with $K = N = 1$, $z = 1$, $Z = 2$ for each ORU. All three alternatives give the same result, equivalent to full cannibalization of ORU 5 and 6.

However, if systems 2 and 3 are combined into one system with $K = 1$, $N = 2$, $z = Z = 1$ for each ORU (similar to the table), the result would not be equivalent to full cannibalization of ORU 5 and 6. Such a configuration implies that there are two pairings of ORU 1 and ORU 2, and at least one specific pair must operate. The redundancy block diagram for this is different from Figure 6-4.

Table 6-3. **Translation of Redundancy Block Diagram (Fig. 6.4)**

System Type	K	N	ORU	z	Z
1	1	2	1	1	1
			2	1	2
			3	1	2
			4	1	1
2	1	2	5	1	1
3	1	2	6	1	1
4	1	2	7	1	1
			8	1	1

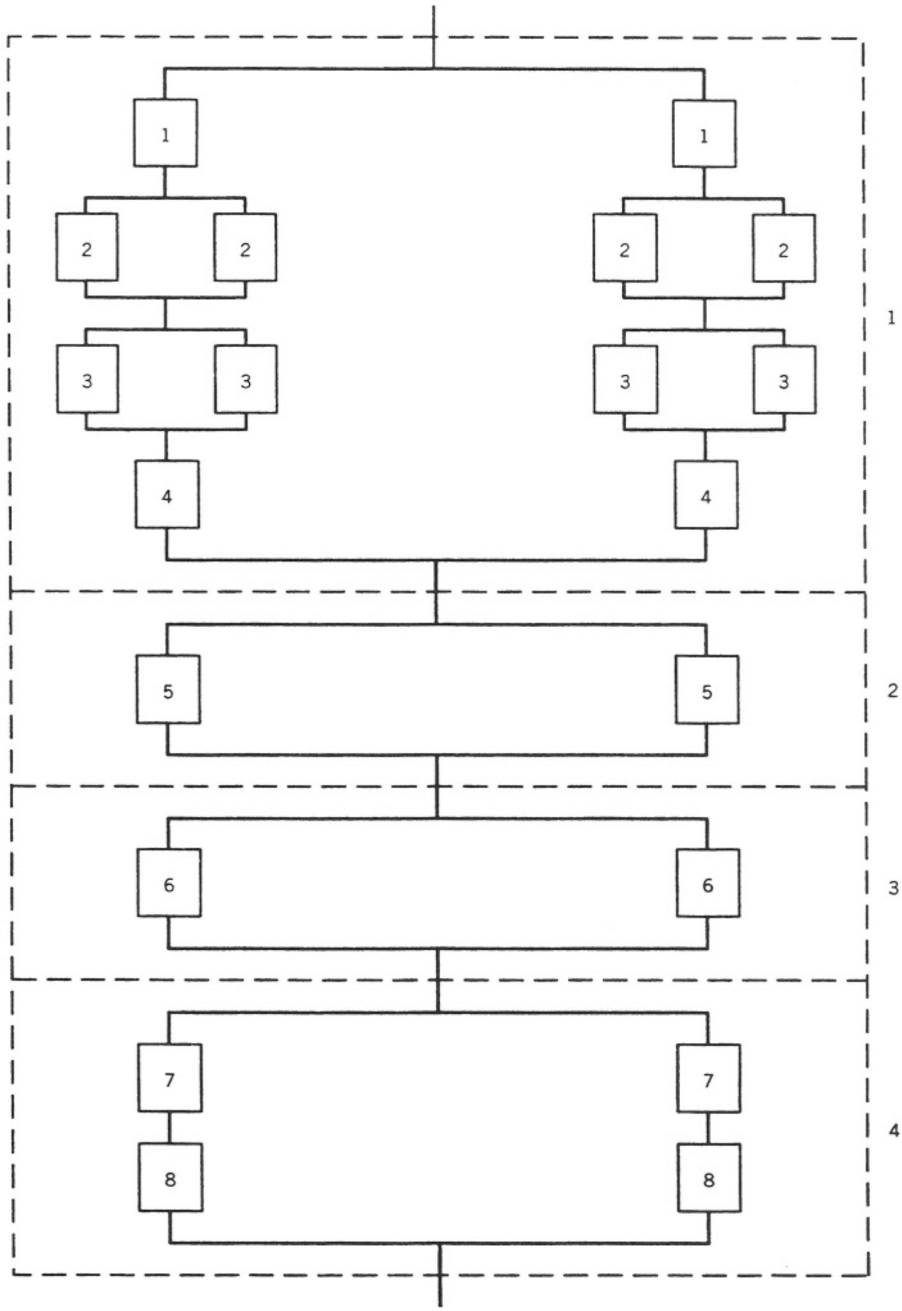


Figure 6-4. Redundancy Block Diagram, space station communications and transmission system. Dashed lines show boundaries of systems

6.13 Numerical Examples

Now we want to provide some real-world, numerical examples from the space station application. Our purpose is threefold: (1) to demonstrate that it is possible to use the theory presented here in efficient models for personal computers that are capable of handling 4000 or more items and up to 10 different system types; (2) to show this theory when compared to traditional approaches produces significant improvements in system performance at any budget level; (3) to illustrate the types of decisions that can be made with the model. The data were taken from the actual design as of late 1990, but do not necessarily represent the final configuration.

Let's examine the traditional approach and identify some of its weaknesses. This approach, used by the contractors until our model was completed, is to buy enough stock on each item so that there is a 95% probability of sufficiency (POS). In other words each contractor was told by NASA to compute levels for on-orbit stock so that on each item there would be no more than a 5% chance of demand exceeding supply during a cycle. As discussed at length in Chapter 1, an item policy such as this has the defect that system cost and performance are fixed outputs. The availability that results may be inadequate or the system cost may exceed the budget. We believe that management should be presented with a system availability-cost curve. Once management has specified a point on this curve, the item decisions are known.

A second defect with the item approach is that resource constraints are not considered. In effect every item with the same demand rate gets the same stock level, regardless of unit cost, weight, or volume. As we have seen in earlier chapters, an important aspect of the system approach is to obtain the maximum effectiveness per resource unit (dollars, weight, volume or whatever). It is possible to use POS as a measure for a system of items, rather than item by item. Thus we could develop an optimal POS-versus-cost curve where the probability of sufficiency is the chance that demand can be satisfied for every item on the system. That would be an improvement over the item approach, but there are other significant defects with POS, as discussed next.

POS was discussed briefly in Section 2.7, where we noted that it is similar to fill rate, except that it is computed from the probability distribution of demand instead of the probability distribution for the number of units in repair. But, these items are repairable, and even though resupply is periodic rather than continuous, our stockage policy should not be independent of the amount of time to repair these items on the ground.

An even more serious defect with POS is that it is being used to compute on-orbit stock only. But the logistician's problem is to determine the total buy of each item, not just the number to put in orbit. The POS calculation assumes that there will always be enough stock on the ground to replenish the orbital stock levels each cycle. It is because the assumption of infinite ground stock is not very reasonable that we developed the theory above to determine the total stock to buy for each item and the optimal split between orbit and ground.

A final defect with POS is that it is insensitive to redundancy. It is necessary to compute the probability distribution of backorders in order to model availability when there is redundancy.

For purposes of comparison we have taken the investment required by the preliminary approach, but allowing for the fact that some units of the item will not be available for the next shuttle. The availability from the contractor policy is evaluated, and that investment is used as a constraint for the optimal model. There were no item volume data, and no target values for the total volume or weight so the only resource considered in these computer runs was cost. The objective is to have 100% power availability or all 8 of the 8 systems operating at the end of a 365 day resupply cycle. (Each system is a series string, and if one ORU fails the system no longer operates).

The input data file to obtain a 95% availability with all 8 of the 8 systems operating was presented in Table 6-2. The system is the PV Module, part of Electrical Power Generation. The 23 items that are dominated by random failure are included in the input. NASA estimated that these can be repaired on the ground within about 160 days, well within the 365 day resupply cycle. There were also 3 items that were dominated by wear out that are not included in this list. They are excluded because it is assumed that when they fail they cannot be repaired. The contractor recommended buys are much larger, and incommensurate with those of repairable items. Of course, the model can include these items as well, but for purposes of comparison with the preliminary POS quantities, we did not want to include them in these runs.

Since the item data relate to the actual program, we have used an ORU number rather than the actual name of the ORU. The data include also the indenture, cost (in thousands of dollars), weight, volume, z, Z, stock level from the preliminary POS computation (IS), variance to mean ratio of demand (l=Poisson demand), average annual demand per unit of the ORU. For example, if Z is 3 and there are 8 systems, the total average annual demand for the ORU, m, is 24 times as large).[1]

[1] Strictly speaking we are modeling the cold-standby case where back-up systems are turned on only when a system fails. Thus, the demand rate stays approximately constant (until there are no more backup systems and an additional system fails). For purposes of data entry we assume that the total item demand rate is divided by N times Z to obtain the demand rate per occurrence of the ORU. The warm-standby case is probably more

Figure 6-5 shows the availability of 100% power for the POS stockage policy and for our optimal policy over the course of the logistics cycle. As noted in Section 6.4, each curve decreases over time to a minimum just before the next resupply shuttle arrives. Note that both curves are slightly convex when viewed from the horizontal axis, because the availabilities are so low. The optimal policy gives higher availability than the contractor policy at every time point, and the average availability of 39.6% is almost three times as large. The item stock levels are shown as Cases 2 and 1 respectively in Table 6-5 and discussed below.

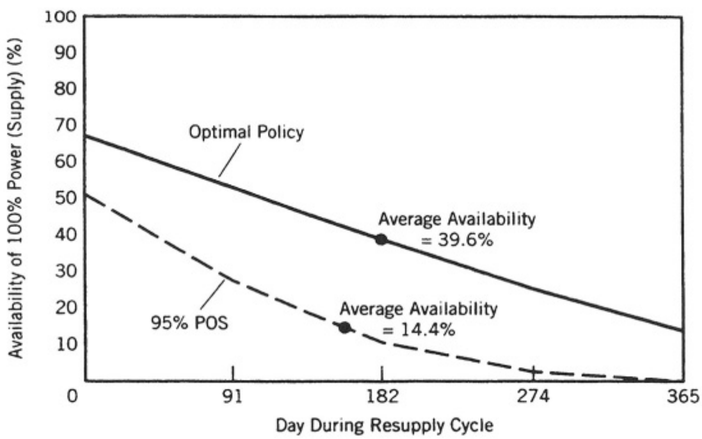


Figure 6-5. **Power generation system, comparison of stockage policy (dashed line) and optimal policy (solid line).**

In order to consider 75% power, we modify Table 6-2 so that $K = 6$ in the first line of data. This is like having redundancy at the system level, but not at the ORU level. In some cases there is a capability to switch/cannibalize ORUs between systems (e.g. in power distribution there is some "cross-strapping" which allows for electrical switching of ORUs between strings without requiring physical cannibalization, but obtaining the same benefit), and this will result in a higher availability for any set of stock levels. If cannibalization of all ORUs were possible, the input data in Table 6-2 should be modified so that each z and Z for each ORU should be changed from 1,1 to 6,8 (or from 3,3 to 18,24 and from 4,4 to 24,32) and K and N should each be 1. In other words for the PV module to operate, any 6 of the 8 occurrences of an ORU must operate, rather than all ORUs on 6 specific systems. These two cases are called no cannibalization (Case 3) and

realistic. As the number of backorders for an item increases, the demand decreases. In Section 6.14 we discuss research that suggests the differences are minimal when the average annual demand is less than one.

cannibalization (Case 5) for 6 of 8 systems (75% power) in Table 6-4. A similar calculation is performed to obtain 4 of 8 systems (50% power), both under no cannibalization and cannibalization in Table 6-4. We show availabilities at the end of the 365 day cycle in each case.

Table 6-4. System Results

Case		Optimal	Cannibalize	Cost($000)	End Cycle Avail. %
1	95% POS	No	No	61,258	0.74
2	8 of 8 Up	Yes	No	61,176	14.41
3	6 of 8 Up	Yes	No	61,176	57.61
4	4 of 8 Up	Yes	No	58,499	86.29
5	6 of 8 Up	Yes	Yes	56,642	87.94
6	4 of 8 Up	Yes	Yes	59,238	99.40

In effect the no-cannibalization case corresponds to redundancy at the system level only, and the cannibalization case corresponds to redundancy at the ORU level only. Of course, there can be redundancy at both the system level and the ORU level. This would be reflected by having K less than N for the system and z less than Z for an ORU. The model has this capability, and the numerical results would fall between the extremes of no cannibalization and full cannibalization shown in Table 6-4.

The input data in Table 6-2 assume that there is no redundancy since $K = N = 8$ and $z = Z$ for each ORU (though the values of Z differ somewhat by ORU). The same case with no redundancy could be specified with $K = N = 1$, where each z and Z are multiplied by 8. This leads to the same availability and optimal stockage policy for the theory that has been developed in this chapter. However, the usual redundancy modeling as in Isaacson et. al (1988) understates the probability that all 8 systems operate as shown in Table 6-2. Instead of the end-of-cycle expected availability of 14.41% that is shown in Case 2 of Table 6-3, that theory yields an estimate of 7.40%. The reasons for this are discussed below in Section 6.14 and Problem 4.

Now we compare the system-level results for various policies in Table 6-4. Table 6-5 shows the item stockage decisions that underlie these. Figure 6-6 shows the computer-generated availability-cost curve for Case 3 beginning at the investment cost of $61.176 million and showing all solution points through a 95% availability.

The major conclusions from Tables 6.4 and 6.5 are as follows

1) The investment of about $61 million suggested by the 95% POS calculation is likely to be inadequate, since only 0.7% of the 365 day cycles are estimated to be completed at 100% power. This is because the contractor model computes only the orbital spares based on the assumption of 95%

availability on each item; infinite ground stock is assumed but not included in their calculation.

2) The solution for Case 5 costs about $4.5 million less than the POS solution. This illustrates the fact that adjoining solutions may have a large cost difference. Here the next item to be procured was a $6.5 million item which would have exceeded the investment constraint that we imposed. Of course, when all systems on the station are combined, the $6.5 million will be small by comparison with the total spares budget and appear much less grainy.

3) Table 6-4 shows that if it is possible to switch/cannibalize ORUs between systems (as in power distribution), the impact on system performance is dramatic (Case 5 vs. Case 3 and Case 6 vs. Case 4).

4) Under the constraint of about $61 million, the optimal stockage policy in Table 6-5 for our model looks significantly different from that of the POS solution. We tend to buy more stock of the higher demand items, particularly if they are lower cost (almost three times as much for ORU #17).

Table 6-5. Stockage Policies

	Cases					
ORU #	1	2	3	4	5	6
1	1	0	0	0	0	0
2	1	2	2	2	1	0
3	2	2	2	2	1	1
4	3	2	2	2	2	2
5	4	7	7	7	7	7
6	8	14	14	15	11	9
7	4	6	6	7	7	7
8	4	9	9	10	8	8
9	7	9	9	8	7	5
10	3	4	4	4	4	4
11	3	4	4	4	3	3
12	4	7	7	8	8	8
13	4	8	8	8	8	8
14	1	2	2	2	2	1
15	2	3	3	3	2	1
16	2	1	1	1	1	0
17	3	8	8	8	7	7
18	1	3	3	3	2	1
19	2	1	1	0	1	2
20	3	4	4	4	4	5
21	3	5	5	5	5	6
22	4	5	5	5	0	0
23	1	3	3	4	3	3

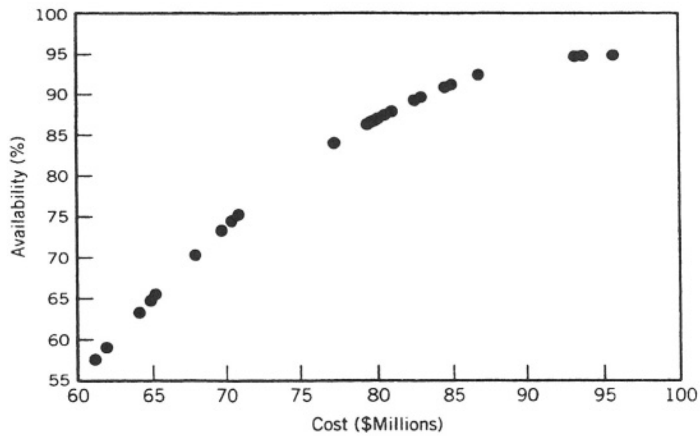


Figure 6-6. **Computer-generated availability-cost curve for 6 of 8 systems up, with no cannibalization**

5) The optimal allocation of a $61 million investment for 8 of 8 systems operating is unchanged when we relax the requirement to 6 of 8, and almost unchanged when we require only 4 of 8 (assuming no capability to cannibalize or switch ORUs between systems). The fact that this stockage policy is "robust" even when the objective changes is highly desirable and encouraging.

6) When cannibalization is allowed, the optimal stockage policy is affected somewhat. The big changes are for three ORUs: #6, #9, and #22. This is because the first two items have a Z of three in each system and the latter has a Z of four. Thus, under cannibalization fewer of these items are needed. However, the cannibalization policies are still quite different from the POS policy.

Table 6-4 showed end-of-cycle availabilities, though average availabilities over the cycle tend to be more meaningful for the electric power system. Figure 6-7 presents data in a format that is more familiar to power system engineers. It shows the expected fraction of each cycle that the system will be at various power levels, assuming no cannibalization. Thus for the optimal stockage policy, the average availability at 100% power is 39.6% of the cycle (as in Figure 6-5); the average availability at 75% power is about 30%, etc. Again we see that the optimal policy yields significantly higher availabilities at every power level, though the spares budget is the same.

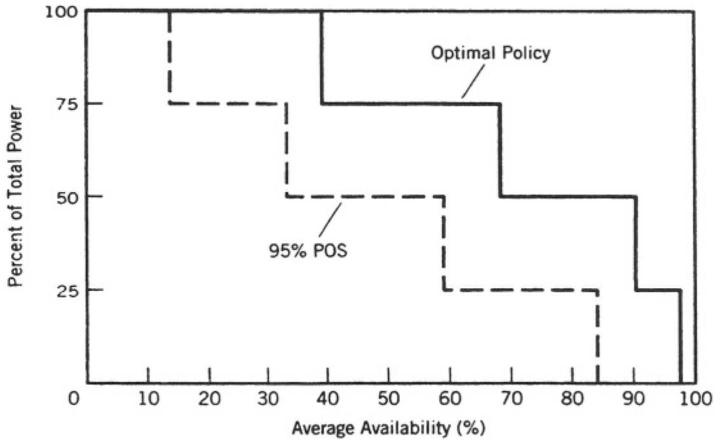


Figure 6-7. **Power generation system ($61.3 million), optimal and 95% POS policy compared**

6.14 Other Redundancy Configurations with 50% ORUs Operating

Now let's examine possible design alternatives available to NASA and how they impact availability and spares cost. In Table 6-4 there were two cases where the Space Station was assumed to operate (be available) if at least 50% of the number of units of each ORU type operates: Case 4 with no cannibalization and Case 6 with complete cannibalization. In fact, there are several other combinations of K of N systems and z of Z ORUs that provide availability if at least 50% of the number of units of each ORU type operate. Each combination represents a different physical design configuration which may cost NASA more or less money to implement on the station. We do not know these design costs, but we can calculate the benefits in terms of spares costs and system availabilities.

Table 6-6. Alternative 50% Power Configurations ($Kz = .5NZ$)

					Spares Cost Target		Availability Target	
Case	K	N	z	Z	Cost($000)	Avail. (%)	Cost($000)	Avail. (%)
a	1	2	4	4	61,027	47.67	105,529	95.58
b	2	4	2	2	61,176	66.48	93,627	96.42
4.c	4	8	1	1	58,499	86.29	70,340	95.34
d	4	4	1	2	60,987	89.65	73,597	95.21
e	2	2	2	4	60,705	97.12	56,726	96.33
6.f	1	1	4	8	59,238	99.40	30,433	95.04

Table 6-6 contains both Case 4 and 6 from Table 6-4 plus the four other configurations that require at least 50% of each ORU. We show explicitly the values of K and N as well as the values of z and Z for the first ORU (the values of z and Z are the same for every ORU except that ORU #6 and ORU #9 are a multiple of 3 for z and Z; ORU #22 is a multiple of 4).

The six alternatives are listed in order of increasing efficiency. As noted earlier, Case 6 corresponds to full cannibalization and produces the highest availability. Case 4, called no cannibalization in Table 6-4, is in the middle of the remaining five cases. It is hard to decide what to call these cases, but the first is clearly the worst way to interconnect ORUs so that the system is available when at least 50% of each ORU operate - to require that all four ORUs of each type be up in one of two systems.

From the physics of the configurations there is an obvious pattern in the way in which values of z, Z, K, and N change from least efficient to most efficient. This is perhaps even clearer from the redundancy block diagrams a-f in Figure 6-8. At first glance it might appear that the availability from lines 3 (Case c) and 6 (Case f) in Table 6-6 should be similar, since any 4 of 8 operating ORUs result in 50% availability. But, that is true only with one ORU. As more ORUs are combined, line 3 requires that the operating ORUs be on exactly the same 4 systems, and this is much more restrictive.

Table 6-6 dramatically illustrates the richness of the computer model in terms of evaluating various redundancy designs. When the system cost target is held constant at no more than \$61.3 million, there is a dramatic range of availabilities from 48% to 99% depending on the type of redundancy. Since it costs more to increase availability when the availability is high, the results are even more dramatic when we hold the availability target constant at 95%. The spares investment cost in the next-to-last column of Table 6-6 has a range of about 3.5 times from the least expensive configuration costing \$30 million to the most expensive at \$106 million.

The more efficient redundancy configurations may cost more to build or they may have additional on-orbit weight and volume implications. Some may not be feasible to build. Nevertheless, it would be extremely desirable for system designers to consider logistics issues such as spares when they are building systems. Now we have a tool that can be used in assessing these tradeoffs.

The change in allocation of the \$61 million between Case 4 and Case 6 was discussed earlier. The policy for the top two cases in Table 6-6 (Cases a and b) is identical with Case 2 in Table 6-5. Thus, the optimal stockage policy is very "robust" with respect to changes in the redundancy structure.

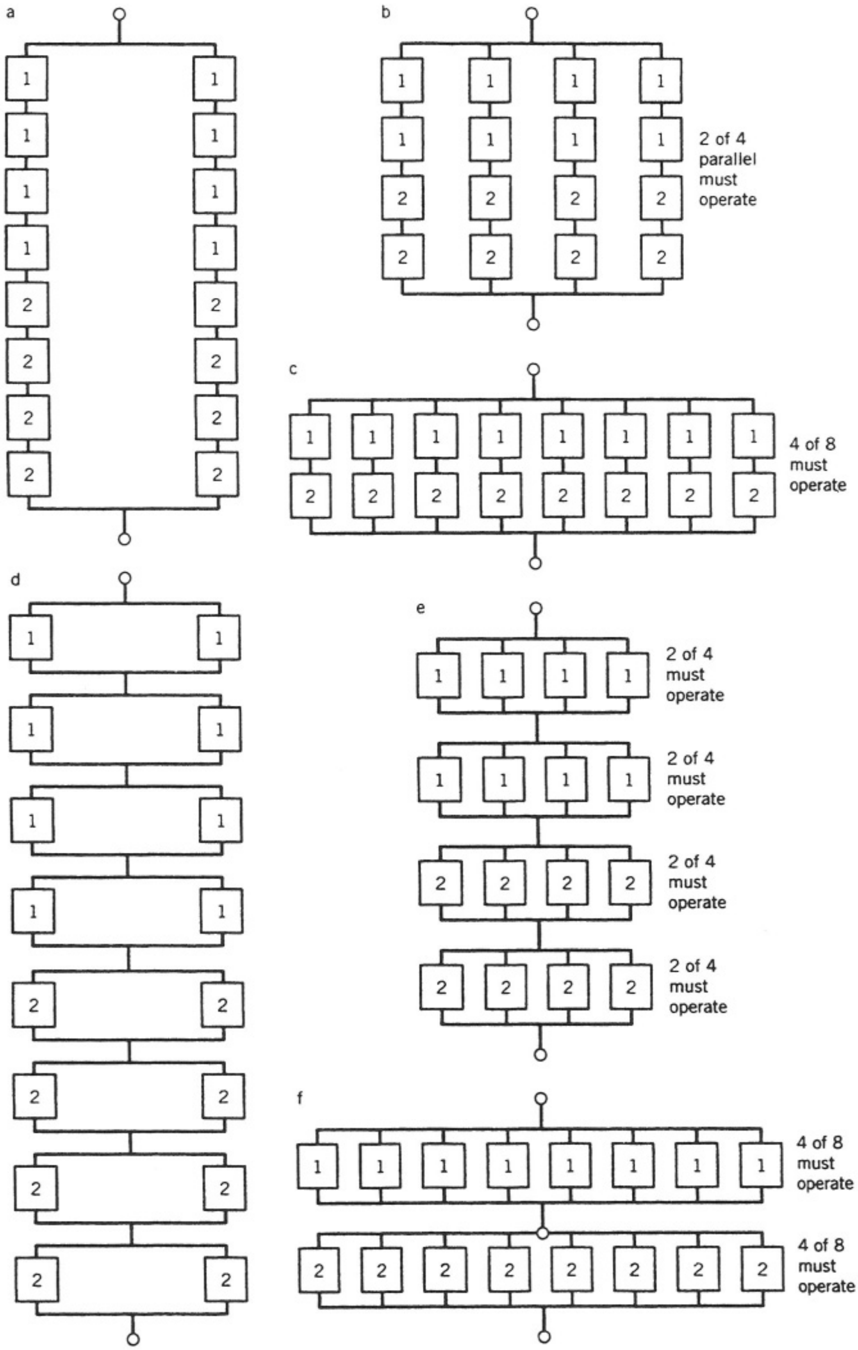


Figure 6-8. **Alternative 50% power configurations**

6.15 Summary of the Theory

Now that the theory development is complete, it is useful to review some of the key assumptions and consider what generalizations are possible. How does this theory differ or complement the theory in previous chapters?

The cornerstone of the repairable item theory in earlier chapters was Palm's theorem which says that the probability distribution for the number of units in repair is independent of the shape of the repair distribution. Palm's theorem was never used in this chapter, because it isn't true here. Equation 6.1 explicitly recognizes that each demand has a probability m_2/m of requiring more than one cycle to repair, a probability m_3/m of requiring more than two cycles, etc. The percentage of repair completed before each shuttle flight is the important factor, not just the mean repair time. Of course, the result is unaffected by the exact time within a cycle that the repair is completed.

On the other hand the spirit of the infinite channel queueing assumption is alive and well here, as in previous chapters, in the sense that there is ample repair capacity and the repair time for an item is not influenced by the number of items already in repair. As before, the time to perform repair is assumed to be dictated by the type of failure and the complexity of the repair.

We have assumed that the cycles are of equal length, but the equations are completely valid if the cycles differ in length. Of course, the computer program and the data manipulation have to be a little more sophisticated. Similarly the demand rates can change over time, as they will during the build up of the station. The theory applies to multi-echelon stockage, and is easily extended to multi-indenture cases (see Problem 7). However, the computer program becomes significantly more complicated, because whenever an SRU stock level increases, the repair delay for the "parent" ORU is reduced, requiring a recomputation of its probability distribution. As in the theory of previous chapters, it can be extended to more echelons or indentures if the "physics" of the problem warrants. Since commonality tends to increase at lower indentures, not just across "parents" in the same system but in different systems, the extra complications are probably not worth the effort.

It should be noted that there has been some modeling of redundancy for the case of continuous resupply rather than periodic and a fleet of aircraft instead of a single station. Isaacson et. al. (1988) allow redundancy at the first indenture item level only. The allocation of item backorders to each aircraft in the fleet is conceptually similar to our problem of allocating ORU backorders to systems on the space station. However, they use the infinite population assumption to distribute item backorders to specific aircraft, and then compute availability of the aircraft fleet by using a binomial distribution.

This simplifies the computation significantly, not only that of $S(i)$ for a particular item from Equation 6.6 above, but that of $S(i)$ for all items in a system and the availability as described in Sections 6.7 and 6.8, respectively. However, the infinite population assumption is unrealistic for most system types on the space station, where the number of identical systems is small. We show in Problem 4 that the simplification produces estimates that are optimistic by about 5% in half of the designs. As the number of systems increases, the infinite population assumption produces a smaller error.

The expected number of systems down is the same for the binomial and the improved approach. But, the variance in the probability distribution for the number of systems down is greater under the improved theory, particularly when N is small. Thus, the binomial theory *understates* the probability that all N systems operate, but *overstates* the probability that at least 50% of the systems operate. The improved theory is particularly important if we need to estimate the probability of no power, in which case the space station must be evacuated. Under alternative b in Figure 6-8 the probability that all 4 systems are down is 5.25% from the binomial theory, but 11.02% from the improved theory.

Kaplan (1989) addresses a problem that is similar to Isaacson except that he does not calculate the probability distribution for the number of operating systems. The modeling at the item level in Kaplan is slightly more general. He does consider downtime for remove-and-replace maintenance (which we include in a different way in Chapter 7), and the reduction in demand for an item that occurs as the number of backorders increases (warm standby case). In the limit when there is a backorder at every location of the item, the demand rate goes to zero (as it should). This is sometimes called the *finite calling population assumption,* not to be confused with the infinite population assumption for the distribution of backorders to aircraft discussed previously.

Kaplan tests the adequacy of his approximation by simulation. The conclusion of most relevance to the space station is that when the mean time between failures on an item exceeds a year, the "finite calling population" correction is not needed. Since most items on the space station should have much longer mean times between failures, we will not incorporate this refinement.

6.16 Application of the Theory

It is important to discuss how the theory can be applied. We want to consider uses of a computer model based on this theory at different times in the development and deployment of a system.

Early in system development item costs, weights, and volumes are known imprecisely, if at all. Even the design is still being changed. Ideally the model could be useful to designers in the evaluation of different redundancy configurations. The complexity of the redundancy is a function of the required availability: more redundancy is cost-effective if high availabilities are needed, other things being equal. Depending on the capabilities for electronic switching between ORUs, we may get much of the benefit of cannibalization.

Even though there may be some capability to switch between broken ORUs to keep a system operating or even cannibalize, logisticians are likely to calculate the spares requirement under the assumption that every "hole" will be filled, regardless of redundancy. If that policy is followed in provisioning, it would be sensible to evaluate the availability under the K of N degraded mode for each system that represents minimum capability. For example, we might provision to 90% availability for all critical systems on the station, but find that those same levels evaluate to a 98% assurance that no critical system will be below its minimum requirements for K operating of its N systems.

Item costs, weights, and volumes are hard to obtain and demand estimates even harder (and more important), but the system cost-availability curves are likely to provide some information. That is, at the system level some of the item data errors will cancel out, and the overall system cost-availability curve should be meaningful. If management looks at these curves for several systems, it should be in a better position to set meaningful availability targets, weight and volume constraints for different systems.

Some major logistics decisions can (and must) be made from this early data, and a model is better than a guess. For example, how long should the cycle between resupply flights be? Are there ORUs that are similar and which could be made common? What is the tradeoff in extra design cost, weight, volume to achieve commonality versus reduction in spares? We will discuss such questions in the next chapter.

As the development matures, more precise data estimates should be obtainable. Eventually the same techniques can be used to calculate optimal shuttle "manifests", i.e. based on the latest failure rate information, age of installed ORUs that are subject to wear out, constraints on the available weight and volume for the next shuttle and in space, what should be sent on the next shuttle?

6.17 Problems

1. Evaluate Equation 6.2 when the distribution of due in, Pr{DI = X} is Poisson with mean .2 and p is Poisson with mean .1, $s_O = 1$, and $s_G = 1$.

2. For the single ORU considered in the example of Table 6-1, determine the probability that only System 1 operates, $S(1)$ and $A(1)$, the availability of one or more systems.

3. Suppose that the example of Table 6-1 is modified so that Z (the quantity of an ORU in a system) = 4. Compute the appropriately modified Table 6-1.

4. Derive the expressions for S(3), S(1), and S(0) for Section 6.9 from Table 6.1. Also verify that the binomial probability distribution for the number of systems down agrees with the numbers shown in Section 6.9.

The availability estimates from the binomial theory as applied to the six lines of Table 6-6 yield the following availability estimates for the \$61.3 million cost constraint: 52.28%, 71.91%, 91.00%, 89.10%, 97.05%, 99.40%. Note that the first three of these binomial estimates overestimate availability by about 5%, whereas the last three are quite close.

5. In order to estimate the average availability for a specified stockage policy, it is necessary to calculate $A(i)$ at several time points t in a cycle. If the availability at t = 0 starts very high (say 95%), it will decrease slowly at first and more quickly as t nears 1. Thus, it is said to be concave when viewed from the x axis. If the availability at t = 0 starts quite low (say 10%), it will be convex decreasing more rapidly at first and slowing as t nears T.

The accuracy of the estimated average will increase with the number of points t at which the availability is computed. Suppose that the average is to be estimated from three points in the cycle. What are the best values of t to choose? What is the generalization of this to an arbitrary number of points? Does the choice of points change if the availability values are weighted unequally (e.g. Simpson's Rule in approximating integrals)?

6. When Equation 6.6 is used to calculate $S(i)$, it is not necessary to compute Pr{BO = y} for all y. Show that y less than or equal to Z-z can never result in a system down, and thus is a lower bound; that $(N - i)Z + i(Z-z)$ is the upper bound for y at which there is some probability of i or more systems operating.

7. Extend Equation 6.1 for the number of items in repair to include the multi-indenture case. Assume that when the ORU is examined on the ground there is a probability q_1 that the failure is due to SRU 1 and a probability q_2 that it is due to SRU 2. The ORU can be repaired within one cycle if the SRUs are available, but the SRUs have a probability of requiring one or two periods to repair. Does it complicate the model if those items requiring two periods to repair are fixed by the contractor?

8. Extend Equations 6.7 to the case of N - 3 and N - 4 systems up. (Hint: the sum of the coefficients in the three equations shown are 1, 3, and 9. The sums for N - 3 and N - 4 are 27 and 81. Use a tabular form where the rows are the number of systems up for ORU 1 starting with N, N - 1, N - 2 . . . and the columns correspond to ORU 2. The table entries are the multiplicative coefficients for the probabilities. The advantage of the table is that it highlights the symmetry).

9. Design a computer algorithm that will generate the cases for the hypergeometric distribution of Equation 6.5 efficiently for any N, Z and y. For example, suppose that there are eight backorders to be distributed across three systems with a Z = 4. The algorithm should generate the cases 440, 431, 422, 332 and use combinatorial factors for symmetric cases.

10. There are some redundancy designs that cannot be modeled precisely with the K of N systems and z of Z ORUs operating. Consider the system shown in Figure 6-9, where current can flow on item 3 only in the direction of the arrow, and show that it is not possible to represent this with a redundancy block diagram of the required form. Find a way to model a lower bound on availability (by assuming that item 3 is always broken) and an upper bound on availability (item 3 is always functioning and current can flow in both directions). Note: This type of redundancy is unlikely at the ORU level, though it could occur at the SRU level.

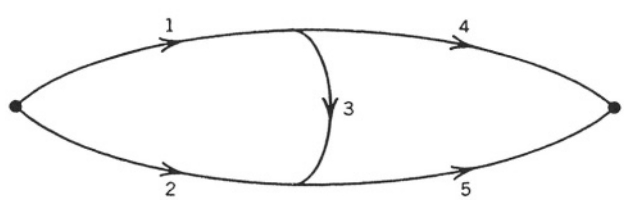


Figure 6-9. **Diagram of redundancy design for Problem 10. Arrows indicate direction of current flow.**

11. (Research) It was noted in Section 6.15 that Kaplan used a simulation and found the approximation for a "finite calling population" was not required if the mean time between failures exceeds a year. However, he assumed that backorders could be assigned to systems using a different infinite population assumption that we have shown leads to inferior estimates of availability. The research problem is to write a simulation for a single ORU and compare the results with our redundancy theory model ignoring the "finite calling population" correction of Kaplan. Find the parameters that define any region of poor agreement, and see whether the suggested correction of Kaplan makes a significant improvement. The

parameters may include the mean time between failures, the time required to repair the item, and z, Z, K, or N.

Chapter 7

SPECIAL TOPICS IN PERIODIC SUPPLY

Sometimes I think the surest sign that intelligent life exists elsewhere in the universe is that none of it has tried to contact us.

- Bill Watterson, "Calvin and Hobbes"

7.1 Chapter Overview

The purpose of this chapter is to present several enhancements to the theory in Chapter 6. These are special topics that are important, but which would have been a distraction from the primary theoretical development. This is an opportunity to provide some further examples of analysis used for management decisions on the space station as well.

We begin with more discussion of availability, and some adjustments that need to be made when comparing resupply cycles of different length. Then we show how the degradation to availability due to maintenance remove-and-replace time in orbit can be incorporated into an overall availability measurement.

Next we discuss a major topic, modeling of items whose demands are primarily due to wear out. There are several such items in the space station's electrical power system, including batteries and solar arrays. The reliability engineer will typically estimate a failure rate (i.e. demand rate) that increases when the item approaches its expected lifetime. But to use our spares modeling techniques, the failure rate curve must be converted to a probability distribution for the time to failure. We show how it is possible to approximate any failure rate curve to any degree of accuracy, and compute analytically the distribution for time to failure.

We will find that it is desirable to keep track of the time at which each item with wear-out characteristics is installed on the station, because this provides a more accurate basis to estimate spares requirements over the next cycle. With the installation dates of wear-out items and knowledge of the spares on-orbit by condition, broken and serviceable, we can prioritize what should be sent on the next shuttle resupply flight. This is sometimes referred to as the *shuttle manifest problem.*

For the purpose of procuring spares that will be delivered many months in the future, the ages of the installed items at the beginning of a resupply cycle are unknown. This simplifies the optimization problem, and allows us to use an analytic approach based on the assumption of a gamma distribution for the time to failure. We are able to calculate analytically the probability distribution for the number of demands during any period of time, even though the original failure rate changes with time.

Finally there is a discussion of how the model can be adapted to a more dynamic environment, as will be encountered during the station assembly phase. It is desirable to link the spares requirements to the procurement actions that must take place at least a lead time earlier and at a time when production lines are still operating.

7.2 Availability over Different Cycle Lengths

In the previous chapter we noted that end-of-cycle availability may be a more appropriate focus for some systems such as the Environmental Control and Life-Support System (ECLSS), that relate to safety of flight. For other systems such as electric power, the average availability (at various power levels) may be more meaningful. We have shown how to calculate both.

However, if we are comparing end-of-cycle availabilities for cycles of different length, we must make an adjustment. For example, an end-of-cycle availability of 90% for an annual resupply cycle implies that over the thirty year program, it is expected that the system will fail about three times since (1 - .9)(30) = 3. By contrast, if there are N cycles per year, a 90% end-of-cycle availability implies that over the thirty year program the system is expected to fail about $3N$ times.

The correction factor to convert the end-of-cycle availability for any cycle length to an annual quantity is just a product of reliabilities:

$$\text{Annual end-of-cycle availability}/100 = (\text{end-of-cycle availability}/100)^{N} \tag{7.1}$$

where *N* is the number of cycles per year. *N* need not be integral and it may be less than one. Thus, if the number of cycles per year, *N,* is four, an end-of-cycle availability of 90% becomes only 65.6% annualized.

Figure 7-1 shows a comparison of annualized availabilities for different resupply cycle lengths and a spares investment of $61 million for the power generation system. This is the same case that we considered in Chapter 6, and the end-of-cycle availability (annualized) for 365 days is 14.4% for the optimal policy and 0.7% for the 95% POS policy as in Table 6.4. While the end-of-cycle availabilities for 100 percent power in Figure 7-1 are all quite low, the shorter cycles lead to much higher availability. It is clear that a semi-annual cycle is much better than a longer period, and not much worse than a 91 day cycle. The reason for the latter is the 160 day ground repair time assumed for the ORUs; even if there is a shuttle every 90 days, it is assumed that nothing can be repaired that quickly. On the basis of this analysis, NASA decided to change from a 365 day cycle to a 182 day cycle for these spares.

When average availabilities are used, instead of end-of-cycle availabilities, no correction factor is needed to annualize availability.

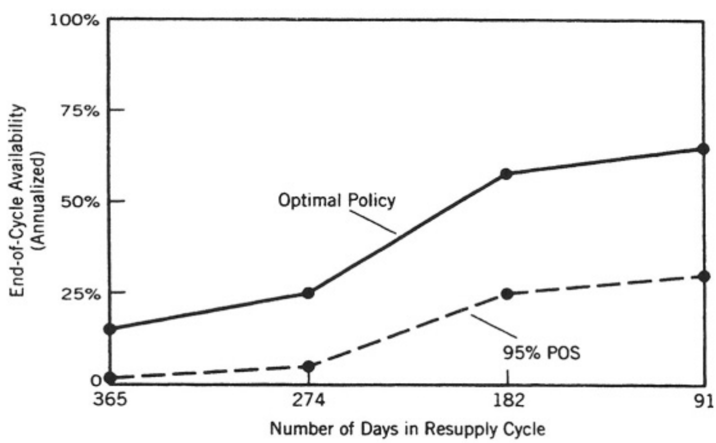


Figure 7-1. **Power generation system, sensitivity of availability of 100% power to resupply interval with $61.3 million spares. Comparison of optimal and 95% POS policy.**

7.3 Availability Degradation due to Remove/Replace in Orbit

The availabilities that we have considered in this chapter and its predecessor are concerned with supply availability only. Thus, if there is enough stock in orbit, it is possible for the availability to approach 100%.

As in Section 5.16, for continuous resupply, we want to take into account the degradation in availability due to the time required to remove and replace ORUs in orbit. Thus, the operational availability is the product of availability due to supply shortages and availability due to maintenance remove-and-replace actions.

To calculate the availability due to maintenance actions, it is necessary to postulate a maintenance policy. The policy for the space station depends on the ORU that has failed. If the ORU failure can be replaced inside the space station and a spare is available, we assume that the maintenance will take place with minimal delay. Thus the fraction of time that ORU i is down for remove and replace maintenance, ρ_i, is just the demand rate times the average repair time:

$$\rho_i = \frac{\text{(aver. demands/cycle)(aver. maintenance hours/demand)}}{\text{hours/cycle}} \tag{7.2}$$

Some ORU failures require that a maintenance action be performed outside the pressurized module (an external vehicle action or EVA event). It is necessary for two astronauts to put on space suits, and their time outside the capsule is strictly limited. The number of these EVA events is planned to be about one per week.

Thus the down time for a remove and replace on an ORU requiring EVA time is the time to the next EVA mission plus the maintenance time. Assuming that all failures requiring EVA for which there is a spare ORU in orbit can be replaced during the next EVA event, Equation 7.2 can be used, but the average maintenance hours for these ORUs must be increased by half of the time between EVA events.[1]

The availability due to remove and replace maintenance for all ORUs on a system is:

$$A = 100[1 - \rho_1][1 - \rho_2]\ldots[1 - \rho_I] \tag{7.3}$$

which is identical with Equation (5.38) for the continuous resupply case. The operational availability due to both supply and maintenance delays is just the product of the two availabilities, Equation 6.8 and Equation 7.3, respectively (divided by 100 since this factor appears in both equations). Note that the availability due to maintenance is constant over the cycle; it does not

[1] Since the electrical power ORUs must be replaced during EVAs, there is an average waiting time of one-half week before remove-and-replace maintenance can be performed. As a result, the maintenance availability for the 26 electrical power ORUs was only 0.954.

decrease over the cycle similar to the availability due to supply. Thus, its impact is the same for end-of-cycle or average availability.

This is strictly correct when the number of systems required to be up, K, equals the total number, N. When K is less than N, the availability in Equation 7.3 is understated because some failed ORUs that have not been replaced will not reduce the number of operating systems below the K allowed.

A more precise model would have to compute the interaction between supply and maintenance for K, $K + 1, \ldots N$ and use Equation 6.7. We believe this refinement is not warranted in the problem context here. We do need the supply availability as calculated in Equation 6.8 when K may be less than N in order to determine the optimal spares mix (assuming that procurement is willing to plan on having "holes" where there is redundancy, which is probably unlikely), and it is useful to have an estimate of the availability reduction due to maintenance as in Equation 7.3. But, we will not usually multiply them together for the reason just stated.

7.4 Failures due to Wear Out

In the previous chapter we were concerned with ORUs whose demands are primarily due to random causes and having a constant demand rate. In this section we want to consider items whose demands are usually due to wear out. For these items the failure rate will vary with time. There may be a short period of "infant mortality" with a high failure rate followed by a low failure rate, which gradually increases as we approach the design life when the item typically wears out. The solid line in Figure 7-2 is an example of an item whose failure rate varies with time in accordance with a "bathtub" curve. Space station items whose failures are mostly due to wear out include batteries and solar arrays.

We assume that it is possible to estimate the changing failure rate for a wear-out item. But, as in each previous chapter, our models require the probability distribution for the number of demands in any fixed period of time. This requires a two-step process developed below: 1) the failure rates must be converted into a probability distribution for the time to the next failure; 2) the probability distribution for time to failure must be converted into the probability distribution for the number of demands over any specified time period (the state probabilities).

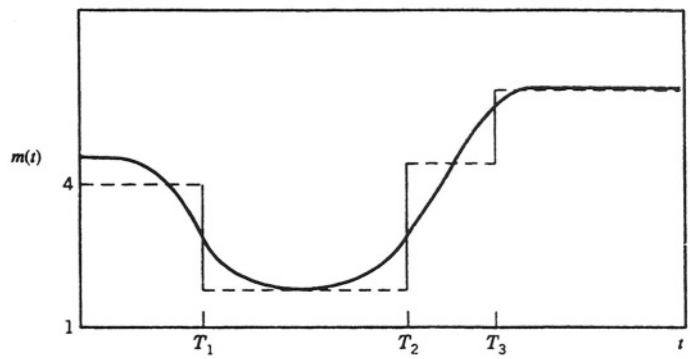


Figure 7-2. Failure rate for a wear-out item. Solid line, example of "bathtub curve." Dashed lines, approximations of failure rates over different time intervals.

This was easy for a constant failure rate. We showed in Section 4.2 that this produced an exponential distribution for the time to the next failure, and a Poisson distribution for the number of demands during any time period. It is important to note that the changing failure rate problem in this section is different from the case of random failures with a mean demand rate that changes over time, discussed at length in Chapter 4. In the latter problem we do not assume that we can predict the change in the failure rate. All we can do is to estimate the probability distribution for the number of demands during a time period from empirically derived formulas.

The relationship between the failure rate $m(t)$ at any time t, and an arbitrary probability distribution for the time to failure, $h(t)$, is a generalization of Equation 4.4:

$$m(t) = h(t)/[1 - H(t)] \qquad t \geq 0 \tag{7.4}$$

where we have written $m(t)$ to indicate that this failure rate is not constant, but may vary with t. The failure rate at time t is the probability density that there is a failure between t and $t + dt$ given that there has been no failure during the interval 0 to t, $[1 - H(t)]$.

For an arbitrary failure rate function $m(t)$, such as Figure 7-2, it is necessary to estimate $h(t)$ using numerical techniques on a computer. But, if $m(t)$ is approximated by a series of constant failure rates over different time intervals, as indicated by the dashed lines in Figure 7-2, an analytic solution for $h(t)$ is possible. The failure time distribution is exponential over each interval where the failure rate is constant, but we must make some scale adjustments.

Suppose we designate the failure rates as m_1 for $0 \le t \le T_1$, m_2 for $T_1 \le t \le T_2$, etc. The probability distribution for time to failure in the first interval is independent of the failure rate beyond T_1 and is simply

$$h(t) = m_1 e^{-m_1 t} \qquad 0 \le t \le T_1 \tag{7.5}$$

The probability of a failure between 0 and T_1 is the integral of Equation 7.5, which we denote as $H(T_1) = a$ for convenience. The probability distribution for time to failure in the second interval is exponential as well, but it must be multiplied by $(1 - a)$ so that when $h(t)$ is integrated from 0 to ∞, the result equals one:

$$h(t) = (1-a)m_2 e^{-m_2 t} \qquad T_1 \le t \le \infty \tag{7.6}$$

The failure rate for $t \ge T_1$ is obtained by substituting Equation 7.6 into Equation 7.4:

$$\begin{aligned} m(t) &= h(t)/[1 - H(t)] \\ &= (1-a)m_2 e^{-m_2(t-T_1)} /[1-(1-a)(1-e^{-m_2(t-T_1)})-a] \\ &= m_2 \qquad T_1 \le t \le \infty \end{aligned} \tag{7.7}$$

Thus we have shown how the constant failure rates are related to exponential distributions of the time to failure. This procedure can be applied to each interval sequentially to yield Figure 7-3. The multiplicative constant for each exponential function is $1 - a$ where a is the area under all of the previous exponential curves (see Problems 2 and 3).

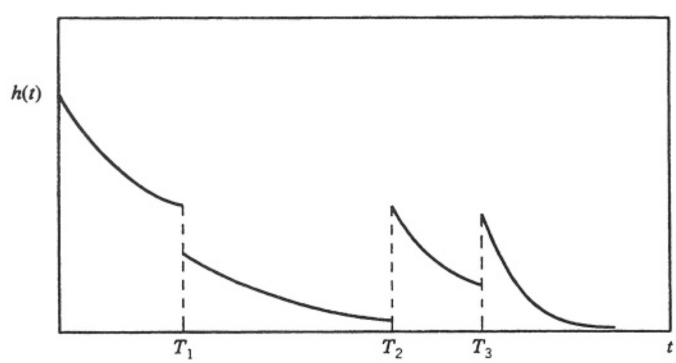


Figure 7-3. **Probability distribution of time to failure for a wear-out item**

The probability distribution of time to failure in Figure 7-3 has a discontinuity at each point where the failure rate changes. Of course, the discontinuities can be made arbitrarily small by using more failure rates in Figure 7-2. The important point is that we have a simple, analytic procedure for computing the probability distribution of time to failure.

Now we are ready for the second step noted above - calculation of the probability distribution for the number of demands over the next cycle from the probability distribution of time to failure just computed. As in the continuous resupply problem of Chapter 5, we will estimate the mean and variance-to-mean ratio of this distribution. When the failure rate in Equation 7.4 varies with time, the estimated variance-to-mean ratio will be smaller if we take into account the age of each installed item at the beginning of the cycle. Of course, in a steady-state procurement model these ages are unknown.

However, in this chapter on periodic resupply we want to consider not only the steady-state procurement problem, but the optimal policy for each shuttle flight based on the latest information including number of units and location of serviceables and unserviceables (failed units) for each item. The information should include the age of each installed item that is subject to wear out as we show below.

7.5 Numerical Example

Suppose that we have ten installed batteries and we want to compute the mean and variance for the number of demands over the next cycle. We assume that the age of each installed battery is known, and that the failure rate is similar to Figure 7-2. The probability distribution of time until failure for a battery conditional on the fact that it has operated for time t is simply obtained by taking the original distribution, $h(t)$, in Figure 7-3 and dividing by the probability that it has not failed during time t, namely $1 - H(t)$. The new probability distribution is defined for $\tau \geq t$ and is like Figure 7-3, except that each probability is multiplied by $1/[1 - H(t))$ so that the area from t to infinity integrates to one.

As in Chapter 6 we assume that the resupply cycle is of length T. Thus the probability that a battery of age t will fail during a cycle of length T is the integral of the probability function in Figure 7-3 from t to $t + T$, divided by $1 - H(t)$. To simplify the mathematics, we assume that if a battery fails during the cycle, its replacement has a probability near zero of failing during the same cycle. In the next section this assumption is relaxed.

Thus for each battery i, based on its age t_i at the beginning of the cycle, we compute the probability ρ_i that it will fail during the next cycle. Thus, there is a probability $1 - \rho_i$ of no failure. The mean number of failures on battery *i* is ρ_i and the variance is $\rho_i(1 - \rho_i)$. The mean number of failures for all 10 batteries is the sum of the means, and the variance for the total number of failures is the sum of the variances since the item failures are assumed to be independent.

This is summarized In Table 7-1, and a hypothetical example given where the ρ_i's range from .1 to .9. We call this the *tracking case* to indicate that we keep track of the age of each battery; the variance is 1.9.

By contrast, consider the no-tracking case, where all we know is that the average $\rho_i = .5$. Thus, the mean is .5 and the variance is .25 for each battery. The total mean is 5.0 again, but the variance is now 2.5 and the variance-to-mean ratio is .5. In order to obtain the same protection level, the no-tracking case with the higher variance-to-mean ratio will require more spare stock.

Table 7-1. **Variability of Demand: Single Failure, Tracking Case**

	Time			Example	
Item #	Since Failure	Mean	Variance	Mean	Variance
1	t_1	ρ_1	$\rho_1(1 - \rho_1)$	.1	.09
2	t_2	ρ_2	$\rho_2(1 - \rho_2)$	.2	.16
3	t_3	ρ_3	$\rho_3(1 - \rho_3)$	.3	.21
4	t_4	ρ_4	$\rho_4(1 - \rho_4)$	.4	.24
5	t_5	ρ_5	$\rho_5(1 - \rho_5)$	.5	.25
6	t_6	ρ_6	$\rho_6(1 - \rho_6)$	.5	.25
7	t_7	ρ_7	$\rho_7(1 - \rho_7)$	.6	.24
8	t_8	ρ_8	$\rho_8(1 - \rho_8)$	.7	.21
9	t_9	ρ_9	$\rho_9(1 - \rho_9)$	.8	.16
10	t_{10}	ρ_{10}	$\rho_{10}(1 - \rho_{10})$	.9	.09
Total		ρ_i	$\rho_i(1 - \rho_i)$	5.0	1.90
Variance/Mean = .38					

The on-orbit stock required for various protection levels, *k,* was derived in Equation 1.2. Thus, the ratio of stock required for a variance of 2.5 versus that for 1.9 can be expressed as a function of the protection level as:

$$\textbf{Ratio of stock} = (5 + 2.5k)/(5 + 1.9k) \tag{7.8}$$

Even for a protection level, *k,* as large as 10 this ratio is only 1.1. However, the comparison can be made more dramatic if fewer of the batteries are near .5 probability in the tracking case. For example, if 5 have a failure probability of .1 and 5 have a probability of .9, the variance in the tracking case drops from 1.90 to .90 and for a protection level, k, of 10 the ratio of stock in Equation 7.8 is almost 1.5, indicating a requirement for 50% more stock.

7.6 Multiple Wear Out Failures at one Location during a Cycle

The theory in the previous section is adequate provided that the probability is very low that an ORU and its replacement could both fail due to wear out in the same cycle. (Throughout this book we have always allowed multiple failures for demand due to random causes). The probability distribution of time to failure need not be gamma or any other well-known function; numerical methods can be used on any probability distribution to estimate the ρ_i i corresponding to any t_i. The failure rate may have a "bathtub" curve shape to reflect "infant mortality," as in Figure 7-2 and combine any aspects of both random failures and wear out.

However, as the cycles become longer, it may be necessary to consider cases where an ORU and its replacement(s) have probabilities greater than zero of wearing out in the same cycle. For each battery location i we must compute the probability of 0, 1, 2, ... failures during the resupply cycle based on the age of the installed battery at the beginning of the cycle, t_i. For example, there will be two failures in one location during the resupply cycle if the first failure in location i is at a time τ_1 between t_i and $t_i + T$, a second failure of the replacement battery at a time τ_2 between τ_1 and $t_i + T$, and no subsequent failures during the interval. The probabilities must be added up for all possible times τ_1 and τ_2, which will require numerical procedures in the general case.

For purposes of spares procurement the ages of the installed batteries at the beginning of a resupply cycle are unknown, of course. Let's consider an analytic approach to this problem that retains the possibility of multiple failures in a location during a cycle and assumes a random distribution of ages for the installed batteries. Suppose that we approximate Figure 7-3 with a continuous distribution for analytic convenience. Because of the increasing failure rate as the item approaches its wear-out life, the probability distribution for time to failure in Figure 7-3 has a hump in the T_2 to T_3 time period.

Two unimodal analytic distributions that could be chosen for this purpose are the gamma and the Weibull. For a particular mean and variance the two distributions look very similar, as was illustrated in Figure 4.4. However, as discussed in Section 4.17, the gamma has the advantage that it is much easier to determine the parameters for any specified mean and variance. That is precisely the capability we need. The mean and variance of Figure 7-3 are computed analytically, and the gamma parameters (a, b) are estimated from the relations following the gamma definition in Equation 4.21 (see Problem 3).

A second advantage of the gamma distribution is that there is a nice physical analogy that may be meaningful. For integral values of the shape parameter a in Equation 4.21, the gamma distribution is the probability distribution for the ath successive exponential event. In some cases it may be possible to visualize the failure process as a series of internal failures, each of which is exponential, where failure of the ORU occurs only after some number of internal failures. In the queueing theory literature where a is an integer, the gamma is referred to as *the Erlang distribution.*

The most important advantage of the gamma/Erlang distribution is that we can calculate analytically the probability distribution for the number of demands during any period of time. For simplicity we begin with a case where $a = 4$ and derive these Erlang-4 state probabilities. Later we generalize to the gamma distribution with non-integral a.

The mean of the Erlang is denoted mT. We will compute the probabilities by relating the Erlang to a Poisson process with mean = $4mT$. If we observe no demands in this Poisson process, there were no demands in the Erlang. If we observe four Poisson process demands, there must have been one Erlang demand since every fourth Poisson demand is an Erlang demand.

The problem arises when we observe a number of Poisson demands not precisely divisible by four. For example, if we observe one Poisson process demand, it may or may not be an Erlang demand, depending on the counting origin. With a random counting origin, there is a probability ¼ that it is an Erlang demand. For a random origin, the general relationship for the Erlang-4 state probabilities, $\text{erl}(x|mT)$, where m is the average annual demand and T is a specified time period, is:

$$\text{erl}(0 \mid mT) = p(0 \mid 4mT) + .75p(1 \mid 4mT) + .5p(2 \mid 4mT) + .25p(3 \mid 4mT)$$

$$\begin{aligned}\text{erl}(x \mid mT) = {} & .25p(4x - 3 \mid 4mT) + .5p(4x - 2 \mid 4mT) + .75p(4x - 1 \mid 4mT) \\ & + p(4x \mid 4mT) + .75p(4x + 1 \mid 4mT) + .5p(4x + 2 \mid 4mT) \\ & + .25p(4x + 3 \mid 4mT) \qquad x = 1, 2, 3, \ldots \qquad (7.9)\end{aligned}$$

where the p's are Poisson probabilities with mean $4mT$. It is easy to verify that the erl's sum to one, and thus comprise a valid probability distribution. It is also easy to check that the mean of the Erlang probabilities is mT (when each equation for $\text{erl}(x)$ is multiplied by x and summed, the coefficient of each Poisson probability, $p(y)$, is $y/4$). The Erlang variance-to-mean ratio approaches 1/4 for large values of *mT,* but must be computed numerically for small *mT.*

While the physical model for the Erlang is based on the *a*th exponential event where *a* is integral, the state probabilities can be computed for nonintegral *a* as well. This allows us to model any variance-to-mean ratio less than one. The generalization of Equation 7.9 is given below. For computational purposes, it is more useful to provide the first three Erlang probability equations, where α is used to denote the integer less than or equal to *a*:

$$\mathrm{erl}(0 \mid mT) = p(0 \mid amT) + p(1 \mid amT)(a - 1)/a + p(2 \mid amT)(a - 2)/a$$

$$+ \ldots p(\alpha \mid amT)(a - \alpha)/a$$

$$\mathrm{erl}(1 \mid mT) = p(1 \mid amT)/a + p(2 \mid amT)(2/a) \ldots + p(\alpha \mid amT)(\alpha/a)$$

$$+ p(\alpha + 1 \mid amT)(2a - \alpha - 1)/a \ldots$$

$$+ p(2\alpha \mid amT)(2a - 2\alpha)/a$$

$$\mathrm{erl}(2 \mid mT) = p(\alpha + 1 \mid amT)/(1 - a + \alpha)/a \ldots + p(2\alpha \mid amT)(2\alpha - a)/a)$$

$$+ p(2\alpha + 1 \mid amT)(3a - 2\alpha - 1)/a \ldots$$

$$+ p(3\alpha \mid amT)(3a - 3\alpha)/a \qquad (7.10)$$

The general pattern can be inferred easily, noting that the numerators of the successive coefficients in an equation increase by 1 until the next coefficient would exceed one. Then the numerators decrease by one until they would go negative, except that the first of these numerators must be computed as in the following example. In the equation for erl(1 | *mT*) the last term on the first line has a coefficient of α/a. An amount $a - \alpha$, less than one, is used to reach the numerator of *a* at which the coefficient is one, leaving an amount $1-[a - \alpha]$ to subtract from *a* for the next numerator, and this equals $2a - \alpha - 1$.

The notation is the biggest problem in understanding these equations. In fact, the equation for erl(2 | *mT*) as written assumes that $2\alpha - a + 1$ is greater than *a*, and that $3a - 3\alpha - 1$ is less than zero. The computer programs to evaluate these equations are not difficult, however.

For each specific value of *x,* the corresponding Poisson probability, *p*(x), appears in one or two of Equations 7.10, depending on whether *a* is integral. Thus, when all Equations 7.10 are summed, the right-hand side is a sum of Poisson probabilities, which must equal one. Since the Erlang probabilities on the left-hand side sum to one and each is non-negative, the Erlang is a valid probability distribution. When the Erlang probabilities, erl($x \mid mT$), are multiplied by x and summed, each Poisson probability, $p(x)$, has a coefficient x/a, showing that the mean of the Erlang is mT.

Table 7-2 shows the variance-to-mean ratios of demand for various Erlang values a (the Erlang is the same as the gamma pdf with integer a) and mean values of demand over a cycle. When $a = 1$ we have the Poisson whose variance-to-mean ratio equals one. Larger values of a lead to smaller variance-to-mean ratios as computed from Equation 7.10. When a is infinite, the time between demands is constant. There appears to be an anomaly when a is infinite and the expected demand is .25. This result arises because with a random time origin and a constant time between demands that is four times the cycle length, there is a probability of .25 of one demand and .75 of no demand. This is like the earlier case of a binomial variable whose variance-to-mean is $(1 - p) = .75$.

Note that as the expected number of demands becomes large, we know more and more about the expected failure times of all failures except the first which was assumed to have a random origin. Thus, these variance-to-mean ratios are appropriate for the case of tracking, when the time of the last demand in the previous cycle is known. When a particular ORU is in several locations, we add the means and variances for each location.

Figure 7-4 summarizes the relationship between the failure rate, the probability distribution of time to failure, and the probability distribution for the number of demands during an arbitrary period of time, t. Any of the three cases (**a**, **c**, or **e**) implies the same process for random failures with a constant mean; any of the three cases (**b**, **d**, or **f**) implies the same process for wear out.

Table 7-2. **Variance-to-Mean Ratio of Cycle Demand - No Tracking - Multiple Failures**

Expected Number of Demands During Cycle	Erlang Parameter (a) 1	2	4	6	Large
.25	1	.816	.759	.752	.750
1	1	.623	.409	.326	0
2.5	1	.550	.313	.231	0
10	1	.512	.265	.183	0
Large*	1	.500	.250	.166	0

* As demands increase, values converge to tracking case.

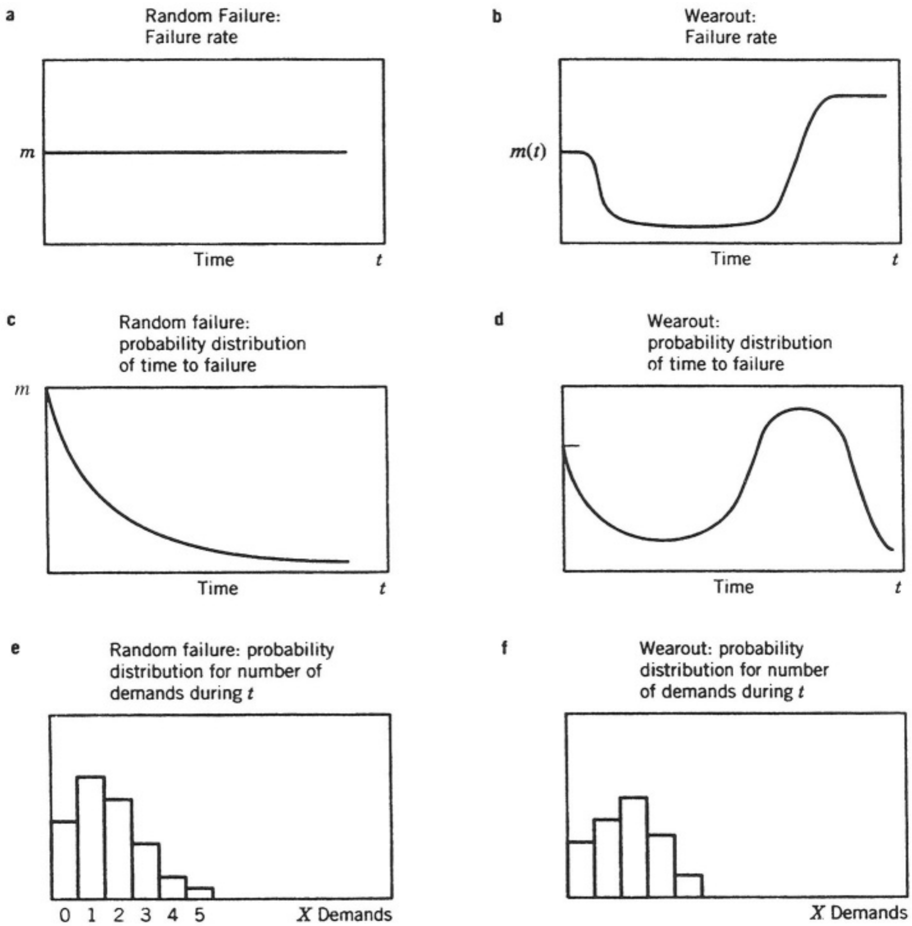


Figure 7-4. Comparison of random failure (**a, c, e**) and wear out (**b, d, f**). Under random failure the failure rate (case **a**) is constant, the time to the next demand (case **c**) is the exponential distribution of Equation 4.1 and the probability distribution for the number of demands during any time t (case **e**) is the Poisson distribution of Equation 2.4.. Under wear out the failure rate typically has a "bathtub" curve (case **b**), the time to failure can be approximated by the gamma distribution of Equation 4.21 (case **d**), and the state probabilities for the number of demands during any time t (case **f**) can be modeled using Equation 7.10 or more easily with the binomial distribution of Equation 4.32.[1]

[1] For random failure items with *changing* means the failure rate of case **a** would be a line that has some variation up and down, the probability distribution of time to failure of case **c** would be a series of exponentials with different means, and the probability distribution for the number of failures during t of case **e** would be the negative binomial of Equation 4.5, but the relationship between these three views is only *approximate,* not exact.

In the wear-out case we showed that if the failure rate is approximated by horizontal line segments as in Figure 7-2, it is easy to calculate analytically the probability distribution of time to failure. The mean and variance of that distribution can be used to fit a gamma distribution, with the advantage that it is possible to calculate analytically the state probabilities for the number of demands during t from Equation 7.10. Again, for simplicity the mean and variance of these state probabilities can be used to fit a binomial distribution.

On the other hand, since the gamma distribution is unimodal, it is not a good approximation to all probability distributions of time to failure. It is possible to consider a mixture of an exponential and a gamma which would be bimodal, but this frustrates the analytic calculation of the state probabilities. Of course, numerical and simulation methods can be used to estimate the state probabilities that result from any assumed failure rate distribution.

7.7 Common Items

The common item problem is more difficult to model correctly than in the continuous review case discussed in Section 5.12 if we are referring to second-indenture items that may be used on multiple ORUs. This is because the importance of an ORU "hole" in the space station depends on where that hole occurs in a system and the type of redundancy at both the ORU and system levels. But the probability distribution for the number of holes on two ORUs (or more) is related in turn to the stockage policy for the common SRU.

Though we will not discuss that complex case, we will consider a simpler case of commonality at the ORU level. In this actual example from the power generation system, there are two types of solar array blankets, a left and right. It is technically possible to have a single, common ORU that could be used in either position, but there are additional design costs and there may be additional weight and other considerations. The data is for the 23 random failure items from Chapter 6 plus 3 wear-out items, of which 2 are the left and right solar arrays.

To evaluate this decision of whether to make the solar array blankets interchangeable, we run the model under both sets of assumptions and calculate both system availability-cost curves. The important observation from Figure 7-5 is that the benefit of commonality increases with the availability target. At about a 73% availability, the dotted line shows a spares budget saving of about \$10 million, whereas at 95% availability the dotted line shows a spares budget saving of about \$15 million. (The actual costs have been replaced by fictitious costs to avoid influencing future contractor estimates for the common item design, but the qualitative results are accurate.)

NASA decided not to make a common blanket because of increased weight; this makes sense if no spare blankets are carried in space, but not if it is necessary to have both a left and right blanket as on-board spares.

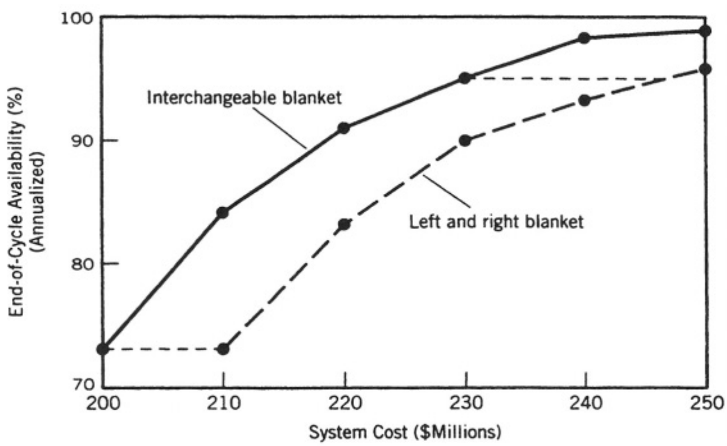


Figure 7-5. **Comparison of cost-availability of having separate ORUs for the left and right solar array blankets or a common, interchangeable blanket. Cost-availability trade-off for all 26 items under optimal stockage conditions with a 365-day resupply cycle and 100% power operation.**

7.8 Condemnations

We have assumed that every ORU that fails in space can be repaired on the ground. In Chapter 6 we outlined the solution for the multi-indenture problem as well where the ground stockage of SRUs required to repair ORUs was determined. Again the assumption was that SRUs can be repaired.

Now we consider the fact that some ORUs or SRUs may not be repairable on the ground, and may have to be scrapped or condemned. Without loss of generality let's consider an ORU which has a probability f that a failure due to either a random cause or wear out can not be repaired.

We need to distinguish two resupply scenarios: (1) one-time procurement; (2) continuing procurement. Many of the ORUs on the space station can only be purchased prior to full-scale deployment while the production lines are still open. This means that in addition to the spares determined by the theory of Chapters 6 and 7, we must buy spares to cover the expected condemnations over the life of the space station, plus some safety stock. It is obviously very difficult to estimate condemnation rates over the next thirty years for a system that has not been fielded yet, and it is hard to give precise guidance.

When continuing procurement is possible, the appropriate modification to the theory is to use a repair time in the model equal to the condemnation fraction f times the average resupply time for procurement plus $(1 - f)$ times the average repair time when the item can be repaired. The procurement lead time should include administrative and other delays.

7.9 Dynamic Calculations

The theory we have developed is for steady-state operations. Of course, the reality is that the space station will be built up over time, and the demand rates will increase with the number of ORUs actually in service. Demand rates will tend to be low, and it will take time for enough data to cause us to revise demand estimates. But, over a 30-year program, all of the data elements will change. Models based on this theory are easy to operate; it is possible to run them at different points in the program build-up. Alternatively, one can calculate only the peak requirements when the station build-up is over.

The same model can be used for shuttle manifests with up-to-date information including the age of items subject to wear out. The advantage is that depending on the available weight and volume for both the shuttle and the station, it is possible to prioritize the spares sent on each shuttle. By looking at the increase in availability per unit resource, weight or volume or both, management can determine whether extra capacity should be used for spares or something else.

It is possible to link the spares required by the model to procurement lead times, and constraints on production, to determine a list of what and when to buy, as well as programmatic budgets required by time period.

7.10 Summary

We have shown that end-of-cycle availabilities over different resupply cycles must be annualized if meaningful comparisons are to be made. Corrections to availability to account for remove-and-replace time of an ORU in orbit have been presented.

For items with wear-out characteristics the failure rate changes with time. Though the failure rate may be the most convenient representation for the equipment specialist who is estimating data, we need to calculate the probability distribution of time to failure for our models. We show that if the failure rate curve is approximated by a series of constant failure rates that differ by time interval, it is possible to calculate the probability distribution of time to failure analytically.

To determine the best mix of items for a shuttle manifest, it is necessary only to compute the probability of failure for each unit over the cycle, taking into account the age of each installed unit. However, for procurement purposes this information is unknown, so that a simpler model based on a gamma distribution of time to failure is proposed.

An example of commonality was presented, and procedures for estimating additional stock due to condemnations described. Finally, there was some discussion of the dynamic problems of stockage due to the fact that the station is built up over time.

7.11 Problems

1. Verify that the mean and variance of the exponential distribution from Equation 4.1 are $1/m$ and $1/m^2$.

2. In the model of Section 7.4 with constant failure rates over different intervals, suppose that $m_1 = 0.25$, $m_2 = 0.05$, $m_3 = 1.0$, $T_1 = 1$ and $T_2 = 4$. Write an analytic expression for the probability distribution of time to failure for any $t \geq 0$.

3. Verify that the mean and variance of the time to failure in Problem 2 are 2.208 and 5.47. Suppose this distribution is approximated by the gamma distribution of Equation 4.21, and estimate the parameters *a* and *b*. (Remember that the Erlang is the special case of a gamma with integer values of *a*). Note that the variance-to-mean of this gamma distribution is larger than that of the exponential. Explain why this occurs.

4. Suppose that the assumptions of problem 2 are modified so that any item which does not fail by $T_3 = 8$ is automatically replaced. What happens to the failure rate graph and the probability distribution of failure. Will the mean and variance of the latter increase or decrease?

Chapter 8

MODELING OF CANNIBALIZATION

Did you hear about the cannibal who went to the psychiatrist because he was fed up with people?

-Eve Titus

8.1 Chapter Overview

In Chapters 2-5 we addressed the problem where repair or resupply could be initiated at any point in time or continuously; in Chapters 6 and 7 we considered the case where resupply is periodic. In the latter case we considered the impact of cannibalization (consolidation of ORU "holes" into the smallest number of systems by remove-and-replace maintenance). Now we want to return to the earlier problem where resupply is continuous and show how cannibalization can be modeled.

In most applications, spares are purchased under the assumption that cannibalization will not be practiced regularly. Of course, there will be situations where maintenance personnel cannibalize a few items to prevent a large portion of the fleet from being grounded, but this does not imply that cannibalization is standard procedure (or cost-free) to be assumed in the optimization. Even if cannibalization is not assumed in the optimization of what to buy, we believe it is desirable to evaluate the increase in availability that could be achieved with those spares if maintenance does cannibalize.

The availability under cannibalization is an upper bound on what could be achieved.

We noted in Section 1.11 that a stockage policy optimized under the assumption of no cannibalization tends to produce a robust stockage policy that performs well regardless of the cannibalization policy actually practiced by maintenance; that stockage optimized under the assumption of cannibalization tends to perform rather badly when maintenance does not cannibalize. Even so we may want to optimize under the assumption of cannibalization because: (1) with newer plug-in systems cannibalization may be easy to perform; (2) it may be desirable to anticipate a wartime scenario, in which cannibalization is more likely to take place.

The cannibalization we consider takes place only at the site (base). It is most likely to occur on first-indenture items, but it may occur at lower indentures as well. The linkage between depot and base stocks is still by means of pipelines as in Chapter 5, but the measure of performance at the base is the probability that there will be y or fewer aircraft down at a random point in time. Feeney and Sherbrooke used this objective function in single-echelon, single-indenture models as early as 1963, and it was surely used prior to that time.

The main reason that availability has not been used as the objective function for optimization is mathematical: availability under cannibalization is not separable into independent item computations. The need to use different objective functions, depending on whether cannibalization is practiced, has been an important conceptual limitation. In this chapter we show that availability can be used as the objective function for the cannibalization problem as well, and that marginal analysis produces near-optimal, multi-echelon, multi-indenture solutions. The computation requires substantially more computer time, but the solutions are slightly better and more credible.

Sometimes it is necessary to model non-stationary demand, such as in a wartime scenario where a planned period of surge flying for several days is followed by a lower planned flying rate for a sustained period of time. Full cannibalization is usually assumed. We describe some of the features of Dyna-METRIC, a model designed more for assessment of specified stock levels rather than determination of optimal levels. Recently the Aircraft Sustainability Model has been designed as a multi-indenture optimization that can use either objective function, availability or the probability that there will be y aircraft or fewer aircraft down.

Then we turn our attention to a model originally called DRIVE - Distribution and Repair in Variable Environments. This model attempts to make optimal decisions based on short-term planning horizons and the *actual* condition and location of items, rather than on the steady-state probabilities that are used in procurement models. For this reason the model can make large improvements in availability. This model is discussed in this chapter, because it is assumed that cannibalization will be practiced at the bases.

8.2 Single Site Model

We assume that demand is Poisson with a constant mean, so the probability distribution for the number of units in repair in steady state is Poisson by Palm's theorem. The probability of having y or fewer aircraft down at a random point in time, $G(y)$, where y is a parameter, equals:

$$G(y) = \prod_i P_i(s_i + Z_i y) \qquad y = 0, 1, 2 \ldots \tag{8.1}$$

and P is the cumulative Poisson distribution of the steady-state probabilities for the number of units in repair and Z_i is the number of units of LRU i on the aircraft. Thus Equation 8.1 indicates that the objective function will be satisfied if there are no more than $s_i + Z_i y$ units in repair of each LRU i, where as before, s_i is the stock level on line-replaceable unit (LRU) i and Z_i is the quantity of LRU i on the aircraft. For a single echelon and indenture, this is easily converted to separable, additive item functions by taking logarithms.

$$\log[G(y)] = \sum_i \log[P_i(s_i + Z_i y)] \qquad y = 0, 1, 2 \ldots \tag{8.2}$$

The cumulative probabilities inside the bracket on the right-hand side of Equation 8.1 are not convex even when the P's are Poisson (this was shown in Problem 4 of Chapter 2 for small values of s when the mean exceeds one). Thus it is a little surprising that the logarithms of these cumulative probabilities are convex when the probabilities on the right-hand side are Poisson, negative binomial or binomial (see Problem 1), so marginal analysis can be used to obtain optimal stock levels.

Because of analytic simplicity, the probability of having y or fewer aircraft down has been used for over 25 years as the objective function for cannibalizable items. But there are several reasons to use expected availability as the objective function

1) We use expected availability as the objective function for non-cannibalizable items. Most systems contain a mix of cannibalizable and noncannibalizable items, and we would like to use the same objective function for both types of items.

2) The probability of y or fewer aircraft down requires a specification of the parameter y. The stockage policy can be quite sensitive to the value of y chosen. For example, when $y = 0$ the range of items stocked will be large; when $y = 1$, the range of items stocked is less, because there is no need to stock low-demand items. Further, the measure gives equal credit to 0, 1, . . y aircraft down, even though the smaller values are preferable; equal penalty to $y + 1$, $y + 2$, . . aircraft down, though the lower values are preferable. Thus, it is has the defect of being an all or nothing measure, similar to fill rate, which ignores the length of time a stockout lasts.

3) Many users misinterpret the probability of having y or fewer aircraft down, confusing it with an availability rate. For example, a .85 probability of y or fewer aircraft down is quite different from an 85% availability. The probability depends on the value of y. A stockage list that produces a .85 probability of 0 aircraft down will typically lead to an availability rate greater than 85%, whereas a .85 probability of y or fewer aircraft down, for large values of y, results in an availability less than 85%.

The ability to optimize the procurement of spares under the assumption of cannibalization is becoming more important, because items on newer systems tend to be easier to cannibalize.[1] There may be nonwartime applications for the military and even some commercial applications where a degree of planned cannibalization is appropriate.

The expected availability under cannibalization, denoted by A_C, is the total number of aircraft, N, minus the expected number down, all divided by the total number of aircraft:

$$A_C = 100[N - g(1) - 2g(2) - 3g(3) \ldots - ng(n)]/N \qquad (8.3)$$

[1] [1]The Air Force has estimated that of the 176 LRUs on the F-16 aircraft dominated by random failure that are carried in war reserve kits, there are about 44 (25%) that are hard to cannibalize. *Hard-to-cannibalize* is defined as over four hours of maintenance time or with significant risk of breakage. For the F-15 there are 237 LRUs of which about 45 (19%) are hard cannibalizations. A recent review has reduced those percentages. Maintenance personnel on the A-10 aircraft told us that there are only ten or so items that they find difficult to cannibalize. One example is a pitch and roll indicator that requires recalibration, and which is easy to damage. The percentage of hard cannibalization items would probably be smaller for commercial aircraft. In an Air Force exercise to simulate combat conditions, we noted that maintenance personnel found it easier to cannibalize some avionics LRUs between aircraft, rather than go to base supply for the item.

where $g(x)$ is the probability density of exactly x aircraft down and n is a positive integer that is large enough such that $G(n)$, the cumulative probability, is approximately one. Since N is the number of end items, $n \leq N$. Rewriting Equation 8.3 with cumulative probabilities yields:

$$A_C = 100\{N + [G(0) - G(1)] + 2[G(1) - G(2)] + \ldots + n[G(n-1) - G(n)]\}/N$$

which is equivalent to:

$$A_C = 100\{N + G(0) + G(1) + G(2) + \ldots + G(n-1) - nG(n)\}/N \tag{8.4}$$

It is easy to evaluate the availability for any set of stock levels using Equation 8.4. The optimization problem is that each $G(y)$ is a multiplicative function of the item stock levels, as shown in Equation 8.1. Thus the availability is not an additive separable function of the item stock levels.

We can still apply marginal analysis, but there is no guarantee that every point on the cost-effectiveness curve will be optimal. It turns out that when marginal analysis is applied, the results are usually optimal (i.e. items are bought in the correct sequence). However, Mike Slay was able to construct a single-indenture case where a non-optimal solution is generated.[1] The first solution of 57, 10, 10 at $7700 shown in Table 8-1 is obtained by marginal analysis, and it lies on the convex hull as indicated by the asterisks. At the next step, for $7800 the marginal analysis selects the policy of 58, 10, 10 rather than 57, 11, 10, because the former availability is larger. For the third step to $7900, the marginal analysis selects the policy of 58, 11, 10, as indicated by the asterisk. It does not consider the policy of 57, 11, 11 which has a higher availability of 97.0618%, because the algorithm had already bought 58 units of the first item in the previous step. Finally at the fourth step, to $8000, the marginal analysis selects 58, 11, 11 which lies on the convex hull.

[1] The three first-indenture items in this case have pipeline mean demands of 54.3, 10 and 10 respectively, variance-to-mean ratios of one, and costs of $100. The availabilities are based on 100 end items, though this type of non-optimality is independent of N.

Table 8-1. Example of Nonoptimal Solution Generated by Marginal Analysis (MA)

Stock Levels	Avail (%)	Cost ($)	Convex	M.A.
57 10 10	96.5568	7700	*	*
57 11 10	96.7979	7800		
58 10 10	96.7980	7800		*
58 11 10	97.0507	7900		*
57 11 11	97.0619	7900		
58 11 11	97.3277	8000	*	*

This example demonstrates that the solutions generated by marginal analysis may not be convex and they may be dominated by points that are not obtainable with marginal analysis (e.g. 57, 11, 11 dominates 58, 11, 10). This isn't surprising since Equation 8.4 is not an additive separable function of the item stock levels. However, it was not easy to find an example where marginal analysis produces a non-optimal solution point. Our experience suggests that these non-optimal points are rare, and that after a non-optimal point, the succeeding solutions obtained with marginal analysis are usually optimal. Even though expected availability with cannibalization can produce some non-optimal solutions, we believe its advantages as an objective function outweigh the disadvantages in most situations.

8.3 Multi-Indenture Model

The extension to two indentures requires that the cumulative Poisson P's derived in Equation 8.1 of the previous section be generalized to represent the probabilities for LRUs either in repair or awaiting SRUs in repair. We begin with a single LRU type and its SRU components $j = 1 \ldots J$. The algorithmic procedure is as follows where the probabilities may be Poisson, negative binomial, or binomial:

1. For the LRU, calculate the set of steady-state probabilities that the number of units in repair is x, Pr{# LRUs in repair = x}.

2. For each SRU, calculate the set of cumulative steady-state probabilities that the number of units in repair is y or fewer, Pr{# SRU j in repair $\leq y$}.

3. For the LRU and its family of SRUs, calculate the set of steady-state probabilities that the number of LRUs awaiting SRUs is y or fewer, Pr{# LRUs of type awaiting SRUs $\leq$ y}.

Pr{# LRUs awaiting SRUs $\leq$ y}

$$= \prod_{j=1}^{J} \Pr\{\text{\# SRU } j \text{ in repair} \leq s_j + Z_j\, y\} \qquad (8.5)$$

Because of cannibalization there will be y or fewer LRUs awaiting SRUs if for each SRU j, the number of SRU j in repair does not exceed the stock level, s_j, plus y times the number of locations for the SRU in the LRU, Z_j.

4. Calculate the probability of y or fewer LRU backorders due to LRUs or SRUs, $\varphi(y)$:

$$\varphi(y) = \sum_{x=0}^{y+s_0} \Pr\{\text{\# LRUs in repair} = x\}$$

$$\times \Pr\{\text{\# LRUs awaiting SRUs} \leq y + s_0 - x\} \qquad y = 0, 1, 2 \ldots \quad (8.6)$$

where s_0 is the LRU stock level. A similar computation is performed for every LRU and its family of SRUs (where we assume in this derivation that a given SRU is not common to more than one LRU type), so that we obtain a set of $\varphi_i(y)$ for each LRU type i. Then the cumulative Poisson probabilities, P_i, in Equation 8.2 are replaced by these $\varphi_i(y)$ to generalize that result to the two-indenture case. A cost-effectiveness curve can be generated for the probability of y or fewer aircraft down using the φ_i from Equation 8.6. However, the $\log[\varphi_i]$ are not additive separable functions of the items. Thus, they are not convex for two or more indentures.

Now let's consider the expected availability objective function for this two-indenture problem.

8.4 Optimization of Availability

As noted in Section 8.2, the expected availability with cannibalization objective function is not separable into individual item optimization problems and is not convex, even for a single indenture. Suppose we use marginal analysis anyway and compare the expected improvement in A_C per dollar invested (Delta) if any stock level is augmented. By any stock level we mean each LRU and each SRU. But, after each point on the cost-availability curve is generated, every Delta must be recalculated because of the nonseparability of the objective function. (Obviously nonseparability is a more serious problem than nonconvexity in optimization). We need to find an efficient way to do this.

Mathematical induction is used. We show that it is possible to find an initial solution. Then we show that having computed any solution, it is possible to calculate the next solution.

We can compute the availability corresponding to the initial stock levels for all LRUs and SRUs, where the levels may be zero or positive. To denote this initial solution, we add a zero subscript to the corresponding set of probabilities of no more than y aircraft down, $\{G_0(y)\}$. Now we want to show how we can proceed from any solution with stock levels on every LRU and SRU and a corresponding set $\{G_0(y)\}$ to the next solution. Our objective is to show how we can use marginal analysis efficiently to find the best LRU to augment (later we will extend this to include SRUs as well). We can compute a new set, $\{G_i(y)\}$, which corresponds to augmenting by one the stock level on LRU i

$$G_i(y) = G_0(y)\varphi_i(y|s_{0i} + 1)/\varphi_i(y|s_{0i}) \qquad y = 0, 1, 2 \ldots \tag{8.7}$$

where $\varphi_i(y|s_{0i})$ is used to denote the cumulative probability of y or less backorders on LRU i or its SRUs given that the LRU stock level is s_{0i}. The SRU stock levels are fixed also, but we have not indicated them in the conditioning since we are not going to let them vary here.

In other words the modifications to the $\{G_0(y)\}$ required to compute $\{G_i(y)\}$, for each LRU i, are very simple. These in turn allow us to compute the availability with one more unit of LRU i from Equation 8.4 and the increase in availability divided by the cost of LRU i, the Delta referred to above. Suppose that we designate by k the LRU with the largest Delta, indicating the next solution point. Now we want to show the modifications necessary to the $\{G_i(y)\}$ that will convert them into a new set $\{G_i'(y)\}$ that can be used to find the next solution point.

$$G_i'(y) = G_0(y)\varphi_i(y|s_{0i} + 1)\varphi_k(y|s_{0k} + 1) \;/\; \varphi_i(y|s_{0i})\, \varphi_k(y|s_{0k})$$

$$G_i'(y) = G_i(y)\ \varphi_k(y|s_{0k} + 1)\ /\ \varphi_k(y|s_{0k}) \qquad y = 0, 1, 2 \ldots \qquad (8.8)$$

for all values of i except $i = k$. The first line of Equation 8.8 is converted to the last line by substituting $G_i(y)$. To calculate $\{G_k'(y)\}$ for the LRU k that was optimal, we must calculate the next step

$$G_k'(y) = G_0(y)\varphi_k(y|s_{0k} + 1)\ \ \varphi_k(y|s_{0k} + 2)\ /\ \varphi_i(y|s_{0i} + 1)\varphi_k(y|s_{0k})$$

$$G_k'(y) = G_k(y)\ \varphi_k(y|s_{0k} + 2)\ /\ \varphi_k(y|s_{0k} + 1) \qquad y = 0, 1, 2 \ldots \qquad (8.9)$$

The point is that if we save the values $\{G_i(y)\}$ for every possible LRU augmentation, it is easy to update them at each solution step, because for a particular value of y we multiply by the same ratio involving the optimal LRU k for all i except $i = k$. For the LRU k that was optimal, it is necessary to compute the $\varphi_k(y|s_{0k} + 2)$ and multiply by a different ratio.

The procedure for augmenting SRUs is precisely analogous, though the notation becomes more cumbersome. It is easy to build a computationally efficient model that is economical in terms of storage such that it is possible to optimally allocate investment across 1000 items or so on a personal computer. Note that steps 1 and 2 of the algorithmic procedure of Section 8.3, computing the probability distributions for the number of LRUs and SRUs in repair, need never be repeated, but the probabilities must be stored. Storage is needed for the $\{G_i\}$ corresponding to any augmentations. For example, if the range on y is 0-49 in $G_i(y)$ for each item i and there are 1000 items (LRUs and SRUs), we would need an array of size 50,000. At 4 bytes of storage per element we need 200K for the G's, and a total of 80K for the LRU and SRU repair probabilities. The entire program for 1000 items will fit easily into 512K of storage.

By keeping the $\{G_i\}$ corresponding to any augmentation in core, the calculation time at each step can be kept reasonable. In summary, the procedure is as follows:

1. Determine that LRU or SRU, k, which produces the maximum increase in availability per item cost using Equation 8.4. The function G_k replaces the function G_0, corresponding to our new availability.

2. Recalculate the function φ_k. If the augmented item was an LRU, we must only calculate Equation 8.6; if an SRU we have to calculate Equations 8.5 and 8.6.

3. For each i except $i = k$, we update the functions G_i using Equation 8.8; for $i = k$, we update the function G_k using Equation 8.9.

4. We repeat steps 1-3 generating points on the cost-availability curve until we reach some specified cost or availability target.

Thus, even though the objective function is not separable, there is an elegant procedure for applying marginal analysis. The deltas are not convex either, and since they are recomputed after each allocation, there is no simple procedure for convexification. It turns out that these are not critical defects, as seen below.

8.5 Comparison of Objective Functions for Cannibalization

In order to evaluate the performance of the marginal analysis algorithm, we took some actual Air Force data on the F-111. We compare the probability of y or fewer aircraft down for several values of y and the expected availability objective in Table 8-2. The data are for two indentures, and we have shown above that marginal analysis can not guarantee an optimal solution for either objective function when there are two indentures.

The maximum availability objective does consistently produce lower numbers of aircraft down than the probability of y or fewer aircraft down for any value of y and any budget level. Table 8-2 shows two such points with budgets of \$48 and \$60 million. However, the maximum availability objective reduces the aircraft down by less than one-half percent over that obtained for the probability of y or fewer aircraft down for the best value of y ($y = 4$ for \$48M and $y = 3$ for \$60M). On the other hand, with the latter objective function one must run several values of y to find the best availability for a given budget.

However, there is another number of interest to logistics planners. What is the cannibalization workload that results from the stockage policy? The expected LRU backorders, shown in the last column of Table 8-2, are the “holes” in aircraft. Since cannibalization is the consolidation of “holes” on the fewest aircraft, the number of “holes” is greater than the number of cannibalization actions. Nevertheless, there is clearly a relationship, and it is desirable to keep both expected LRU backorders low and availability high (see Problem 8).

At small investment levels, the allocation of budget under expected availability looks similar to that for the probability of y or fewer aircraft down with a large value of y. As investment is added, the policy shifts to a smaller value of y, eventually moving to $y = 0$. Thus, the availability criterion effectively chooses y values that decrease toward zero as the investment increases. This is reasonable.

Table 8-2. **Maximum Availability vs. Probability of *y* or Fewer Aircraft Down (F-111 War Reserve Spares Kits - 305 LRUs, 262 SRUs, 24 aircraft); Budgets of $48 million and $60 million**

Target Number Aircraft Down	Budget ($ Millions)	Expected Number Aircraft Down	Expected LRU Back Orders
6	47.660	5.138	255.5
	48.226	5.083	254.8
5	47.779	4.969	212.0
	48.345	4.908	211.3
4	48.186	4.915	172.1
3	48.012	5.111	124.4
2	48.011	5.652	79.6
Max. Avail.	48.122	4.894	174.1
5	60.004	3.733	154.5
4	60.003	3.534	116.7
3	60.074	3.403	80.3
2	60.007	3.567	43.7
Max. Avail.	60.044	3.391	65.8

Note: **Solution preceding $48 million shown in two cases where the last item bought cost $566,500, causing a large jump in marginal analysis solutions.**

The stockage policies obtained with marginal analysis can be improved whenever the stock levels for each SRU, s_j, equal or exceed the number of applications of SRU j on the LRU, Z_j, for all SRUs, provided that the cost of a full set of SRUs equals or exceeds the LRU cost (see Problem 3). This simplifies the marginal analysis search procedure, because it is not necessary to consider policies consisting of a full set of SRUs.

In summary, the maximization of expected availability produces slightly higher availability (fewer aircraft down) than the probability of y or fewer aircraft down for any value of y and any budget level. This is true not only for the data in Table 8-2, but for several other cases we have run. We suspect that the improvement in availability obtained will be small in most data sets, though it is useful to have this new technique as a check.

The biggest advantage of the expected availability objective is conceptual - using the same criterion regardless of cannibalization policy. The availability of a system consisting of some cannibalizable items and some noncannibalizable items can be calculated as the product of the two availabilities, and we know how to optimize the stock levels for each piece.[1]

[1] The overall availability is greater than the product of the availability for cannibalizable LRUs and that for noncannibalizable LRUs, because the cannibalizable "holes" can be moved to aircraft that are down for noncannibalizable LRUs. Also, the overall availability is less than the smaller of the two availabilities, because the number of aircraft down for

Another advantage of the expected availability objective is that it is only necessary to calculate one cost-effectiveness curve, rather than a family of curves as in the case of the probability of y or fewer aircraft down.

On the other hand, the curves for the probability of y or fewer aircraft down can be calculated much faster and more items can be handled in a personal computer. More importantly, with this objective function management can see the tradeoff between availability and a measure related to cannibalization workload. As in most optimization problems where there really is more than one resource, it is possible here to relax the expected availability objective slightly and get significant reductions in cannibalization workload (this is analogous to the relationship between cost and weight in the space station as described in Section 6.11 and depicted in Figure 6.3).

There is an interesting relationship between optimizing with different numbers of aircraft allowed to be down and maximizing availability, demonstrated in Table 8-2. With a budget of $48 million, availability is maximized when the number of aircraft allowed to be down is slightly more than 4 (since the expected LRU backorders are 174.1 and 172.1, respectively); with a budget of $60 million, availability is maximized when the number of aircraft allowed to be down is about 2.5 (although we cannot solve for nonintegral values of aircraft allowed to be down), since the expected backorders of 65.8 are about half way between 80.3 and 43.7. In effect, the procedure for maximizing availability is like the procedure for allowing a certain number of aircraft to be down, where this number decreases as the budget increases.

In Figure 8-1, we see similar results for the F-16 where the data points are generated for various values of y, the number of aircraft allowed to be down. As in Table 8-2 we see that by allowing the number of aircraft down to be slightly larger than its minimum, there is a large reduction in LRU backorders.

---

all LRUs is larger than the number down for cannibalizable (or noncannibalizable) LRUs alone.

If we knew the cumulative probabilities for the number of aircraft down due to noncannibalizable LRUs, we could multiply each cumulative probability by cumulative probabilities for cannibalizable LRUs. The resulting probabilities are the G's to substitute into Equation 8.4 for the overall availability. Slay suggests approximating the former probabilities by a Poisson distribution whose mean is the expected number of aircraft down due to noncannibalizable LRUs. Numerical examples show that this is usually a good approximation. However, in some cases the Poisson understates the variance enough so that the availability with a few noncannibalizable items exceeds the availability under full cannibalization, which is incorrect. The problem disappears when a negative binomial distribution is used as described in Problem 7.

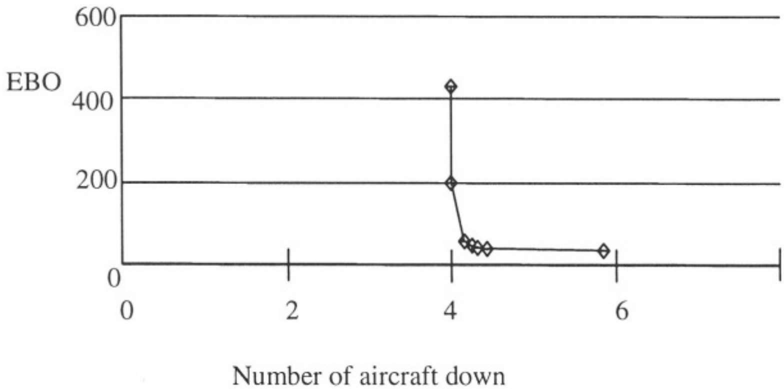


Figure 8-1. **F-16 Tradeoffs of Aircraft Down vs.LRU EBOs ($35.5M)**

8.6 Generalizations

As noted in the derivations of Section 8.3, the probability distributions can be Poisson, negative binomial or binomial. Of course neither objective function is convex for two or more indentures, but we would expect that the marginal analysis solutions under negative binomial or binomial probabilities would lead to results that are not dissimilar qualitatively to what we observed in Table 8-2.

If a multi-echelon model is desired, it is possible to combine the VARI-METRIC theory of Chapter 5 with the theory in this chapter. Cannibalization takes place only at the bases, but the means for the number of LRUs and SRUs in repair at the base must take into account the pipelines to depot which in turn depend on the depot stockage policies.

The theory developed in this chapter can be used to model the case where there is no resupply of the base. Typical military applications are flyaway kits and war reserve kits. That problem is considered below in Section 9.5, where we show that simple changes in the demand rates are the only requirement to use VARI-METRIC theory. Those same modifications to the mean demand rates for LRUs and SRUs in Section 8.3 are appropriate when cannibalization is practiced.

8.7 Dyna-METRIC and the Aircraft Sustainability Model

Dyna-METRIC was developed by Hillestad and Carrillo (1980) as an analytic model for studying the transient behavior of component repair/inventory systems under time-dependent operational demands and logistics decisions like those that might be experienced in wartime. There have been several versions of the model, simulation and analytic, with the most recent analytic version described in Isaacson (1988).

The advantage of Dyna-METRIC is that it is possible to model wartime scenarios such as a planned period of surge flying for several days followed by a lower planned flying rate for a sustained period of time. Cannibalization is normally assumed, though it is possible to assume that some items are not cannibalizable.[1] The mathematical justification is the dynamic form of Palm's theorem proved by Crawford (1981) and described in Appendix A.

Dyna-METRIC is multi-echelon and multi-indenture, but it is best described as an assessment model rather than an optimization model. If the user supplies the stock levels by location, the model will determine the probability that the required number of sorties can be met during the wartime period. In other words Dyna-METRIC computes the probability of y or fewer aircraft down for any set of stock levels. Only in the case of a single base and single-indenture does Dyna-METRIC compute the optimal stock levels.

Slay and King (1987) have developed a model called the Aircraft Sustainability Model (ASM) that incorporates several enhancements to Dyna-METRIC. It allows the user to specify either of two objective functions, availability or the probability of y or fewer aircraft down, and determine the two-indenture, quasi-optimal policy similar to that described above in Sections 8.2 to 8.4. The modeling of the tradeoff between the first- and second-indenture items as described in those sections is more accurate than the procedures used in Dyna-METRIC. Both models are multi-echelon, but Dyna-METRIC does allow the user to specify different base types. ASM was used by the Air Force for the mobility kits in Desert Storm and it continues in use today.

It is important to stress that Dyna-METRIC or ASM is designed for a problem where cannibalization is performed, and when there may be changes in the demand rates because of planned changes to the flying program; the model in Chapter 5 is appropriate when cannibalization is not performed, and we want optimal stock levels in a multi-echelon, multi-

[1] The theory for noncannibalizable items is based on the assumption of an infinite population of aircraft, and as shown in Section 6.15, this can lead to substantial error.

indenture, steady-state problem. But, the VARI-METRIC theory of that chapter is used in ASM to link echelons.

8.8 DRIVE - Distribution and Repair in Variable Environments[1]

In Chapter 7 we noted that better decisions can be made about what to send on a specific space shuttle if the age of each installed unit of any item subject to wear out is known. Now we want to address the corresponding "tactical" problem for continuous resupply. When a procurement model using VARI-METRIC theory is run, a number of simplifying assumptions are usually made. For example, the Air Force assumes that each base is identical - number of aircraft, demand rates, repair capability, order and shipping times, etc. The rationale is that the model is used to determine what to buy, and these procurements do not make an impact until a procurement lead time later, many months in the future. When the procurements are delivered many factors will have changed, and at that time the Air Force can take base differences into account.

This suggests that it is important for the Air Force to have a good technique for allocation of spares - known as *the distribution problem.* Also it is clearly important for the Air Force to have a good technique for depot repair scheduling, since that affects the assets available for distribution.

The author has helped the RAND Corporation in the design and implementation for the Air Force of a model to perform both distribution and depot repair scheduling called DRIVE as described in Abell, Miller, Neuman, and Payne (1992).

8.9 Purpose of DRIVE

Why do we need a DRIVE type model for repair and distribution? As noted above, a procurement model may make a number of simplifying assumptions. But even if the procurement model takes into account all of the base differences, many things will change over a procurement lead time.

[1] Picking a good acronym is both an art and a science. Originally DRIVE denoted *Depot Repair in a Variable Environment.* Then we realized that the model had to deal with distribution as well; fortunately, the acronym meaning could be easily modified.

In Chapter 3 we discussed METRIC. The acronym has been successful in that other modelers have used the term as in MOD-METRIC and DYNA-METRIC. On the other hand, it has led to confusion; after a briefing to Headquarters Air Force Logistics Command in Dayton, Ohio by a RAND colleague of mine, an attendee was heard in the corridor asking why feet and pounds weren't good enough and what did the METRIC system have to do with spares anyway?

This includes item data such as demand rate estimates, unit prices, procurement lead times, etc. and system data such as aircraft basing patterns, flying hour programs, etc.

Suppose that we make the unrealistic assumption that none of those data elements change over the procurement lead time. It is still possible to improve system performance with a good model for depot repair scheduling and distribution. This is because the VARI-METRIC theory is based on the probability distribution for the number of items in repair in *steady-state.* But if we know at each point in time the *actual* number and location of the serviceable and unserviceable units of each item, we can make better short-run repair and distribution decisions.

The percentage improvement obtainable with a good short-run model depends on many factors, but our experience with DRIVE suggests that it is very significant. The Air Force schedules depot repair in most shops over two-week periods (complex repairs for items such as landing gears, engines, etc. will typically have longer repair scheduling periods). The percentage improvement tends to be greater for (a) short repair scheduling periods, (b) when there are more items whose demand rates change with time, (c) when the amount of spare stock in the system is modest in relation to demand rates, etc.

What was used prior to DRIVE? A stock level was computed for each item and location using VARI-METRIC or similar theory. Then bases placed requisitions on the higher echelons that were satisfied on a first-come, first-served basis. Some bases were given a higher priority than others based on the mission, but requisitions were filled on a first-come, first-served basis within priority groups. The problem is that requisitions were based on stock levels, but were unrelated to the amount of stock in the system for each item. If the sum of the stock levels across locations for an item exceeded the actual stock, serviceables plus unserviceables, there would be a number of outstanding requisitions and these would be satisfied in time sequence rather than to maximize aircraft availability; if less, there might be no outstanding requisitions even though distribution to the bases of some depot stock could increase availability.

There are many reasons for the sum of the stock levels on an item to differ from the stock in the system. The stock levels are computed periodically, based on data that on average is a year or two old. Depending on the number of outstanding orders from manufacturers and the number of condemnations for the item, the actual number of spares in the system at any point in time can be more or less than the sum of the stock levels.

The DRIVE model should do a better job of distribution than item managers, because optimal decisions depend on many variables including the tradeoff between an LRU and its SRUs that are not easy without a computer. But, the improvement in repair scheduling appears to be even more significant. To understand the reason, we want to think of the repair requirement in two pieces: a keep-up requirement and a catch-up requirement. The keep-up requirement is essentially the repair necessary to replace failures. This is easy to compute and understand, and historically this has been the primary basis for repair requirements in the Air Force.

The catch-up requirement is the repair required to bring each item to the same readiness level. This requires a model that relates the spare stock by item and location to aircraft availability. Because there was no such model underlying depot repair requirements in the past, the catch-up requirement was not really reflected in the repair schedule. We found in our field tests that this perpetuated the existing problems - an item with a number of base backorders at one point in time tended to remain a problem.

8.10 Model Assumptions with DRIVE

Originally we built a large, combined computer model to make both repair and distribution decisions. Later we realized that this was inappropriate for several reasons. At the beginning of each repair period we need a priority list of what each depot repair shop should attempt to repair, subject to its capacity constraints. And periodically we need to make a computation of where to distribute the assets that have been repaired. However, there are usually significant differences between the list of what we would like repaired, and what is actually repaired.

These lists will differ for many reasons. At the item level this includes lack of repair parts and the inability to find a broken carcass to repair (even though repair scheduling is normally constrained by the number of broken carcasses). At the system level this includes changes in the available repair hours due to the insertion of high-priority repairs of more critical items that are identified during the repair period (or support of a depot rebuild program), test equipment failure or personnel shortages, etc.

Since the model has to be run at different times for repair and for distribution anyway, it makes sense to have two simple models instead of one very complex model. The repair model needs only the detailed information necessary to determine what each shop should repair; the distribution model needs only the detailed information by base that determines where each repaired item should be sent.

Below we list and justify the major assumptions in DRIVE:

1. It is multi-echelon and multi-indenture. The model does depot repair scheduling and distribution to bases taking into account the aircraft availability at each base and the interaction between LRUs and SRUs

2. DRIVE is planning-horizon-oriented. For each base, the model calculates the optimal decisions over a planning horizon. For distribution decisions the planning horizon is an order-and-ship time, because decisions made now will not have an impact until later; for depot repair decisions the planning horizon should be a depot repair cycle plus an order-and-ship time.

3. The model uses repair shop priorities, not a detailed schedule. For each repair shop, the model lists in order of decreasing priority the items to repair during the next repair cycle. A particular item may appear at several points in the list, indicating a need for multiple units. However, the actual sequencing of items for repair during the repair cycle is left to the shop foreman. He is in a better position to take into account the day-to-day factors such as the availability of repairmen and test equipment. When multiple units of an item are to be repaired during a repair cycle, he is likely to find it more efficient to batch them to reduce physical set-up time and repairman inefficiencies.

4. SRU repair parts are assumed to be in stock. When an SRU is scheduled for repair, demands for repair parts may be placed on supply in anticipation of the need. Of course, in some cases the parts needed to repair an SRU are not available. If this is known by the shop foreman before SRU repair begins, it can affect the decisions as to those items to induct for repair. This is another reason why the repair priority list differs from the list of repairs accomplished during the repair cycle.

5. Carcass constraints are considered. Normally the number of broken carcasses at the depot is taken into account in determining the priority list above. In some cases this may be augmented by expected carcass returns to the depot during the repair cycle.

6. Base shortages are consolidated by cannibalization. The Air Force chooses to assume that any "holes" at the base for LRUs are consolidated on the fewest aircraft; any "holes" for SRUs are consolidated on the fewest LRUs. Then depot repair and distribution decisions are made to increase aircraft availability at the bases. In peacetime the Air Force would typically perform only enough cannibalization to meet peacetime availability targets, but the assumption of full cannibalization does provide an upper bound on capability that could be achieved if necessary. Of course, in other applications it may be appropriate to assume no cannibalization.

7. There are timely, accurate data on asset location and condition. This is an obvious, but difficult requirement. If data at different locations is updated at different times during the day, there is a possibility of double-counting or failure to count.

8. There are accurate data on demand rates, repair times, order-and- ship times, LRU/SRU parent-child relationships, etc. It turns out that demand rate errors are much less important to DRIVE than to a procurement model, because we need to project demand only over the planning horizon. This

implies that the problems of changing demand rates and estimates of variance-to-mean ratios are largely unimportant to DRIVE.

9. Each item has one primary repair shop. The basic logic of the DRIVE models constructed so far is that it is possible to designate the repair shop for each failed item. While a given repair may actually involve more than one shop, a more sophisticated model would be needed to model this interaction if a significant amount of multiple shop repair occurs.

10. There is capability for redistribution between bases. This is an optional capability that we built into DRIVE. On the one hand, redistribution should be less necessary with DRIVE, because the system is partially self-correcting based on the latest information. But if the demand rate declines dramatically at a base that has spare stock, the spares will stay there indefinitely unless there is a redistribution capability.

8.11 Implementation Problems with DRIVE

There are a number of practical implementation problems with a model like DRIVE. Some of them are endemic to any model implementation such as user resistance and training problems. Others relate to the adequacy of the assumptions above. But there are some other fundamental concerns that we address below.

1. Master stock numbers versus limited interchangeables. The model deals with master stock numbers, but in the real world an item may actually be several stock numbers with limited interchangeability. That is, a particular stock number may be applicable to only certain aircraft. The depot may have to depend on the requisition submitted by the base to know which specific stock numbers are applicable.

2. Pull versus push system. A system that distributes stock in response to requisitions is called *a pull system;* one that sends stock to bases using central information on asset location and condition is called *a push system.* If the depot were in possession of all information necessary, DRIVE could push assets to the bases. But, as we have just seen in the previous paragraph, the depot is unlikely to have the necessary information on interchangeable and substitute items and the aircraft configurations at the base. As a result the current Air Force implementation of DRIVE is still a pull system where an item is distributed to a base only if there is an outstanding requisition. In effect, DRIVE prioritizes requisitions.

3. Loss of management information and control. Before DRIVE an item manager knew the projected repairs by item over the next quarter, and he knew approximately what would be repaired in the next two weeks. Since the manager controlled distribution as well, he could estimate a delivery date

of a particular item to a specific base. There is still some quarterly planning under DRIVE, but the specific list of items to repair in the next repair cycle is much more dynamic, depending on a number of factors. Furthermore, since the distribution decisions are made by DRIVE (perhaps with some item manager authority to override), the item manager has lost some control and visibility. We believe the gains from DRIVE in terms of better decisions (particularly in terms of the multi-indenture decisions on both LRUs and their SRUs) more than offset the losses, but it is critical to consider the impact of advanced stockage models on logistics managers.

4. Demand from non-aircraft sources. The biggest problem with DRIVE has been dealing with demands other than those from bases flying aircraft. This includes support to depot overhaul programs, foreign military sales, and some contractor repair programs. The DRIVE program managers have resorted to assigning some rather arbitrary priorities to combine these demands into the priority systems for repair and distribution,

8.12 Distribution Algorithm for DRIVE

We want to describe the distribution algorithm briefly. At each base there is some target at the end of the planning horizon: either an allowable number of aircraft down or an availability objective. An allocation of an item of stock to a base will increase the probability for the former objective or increase the expected availability for the latter objective.

The distribution algorithm would be run periodically, say weekly or bi-weekly. It is assumed that any serviceable item already in transit to a base will arrive before the end of the planning horizon. Thus it is counted as if it has arrived in the calculations to follow. Similarly it is assumed that any base repairable item will be fixed by the end of the planning horizon if spares are available at the base.

For the moment we assume the objective function is some target number of allowable aircraft down at the end of the planning horizon for each base, y_j. Later we consider an expected availability target for each base.

Let's consider a single LRU which is sometimes base repairable and sometimes depot repairable and its family of SRUs. Since we are assuming up-to-date information at the depot, we know how many units of the LRU are broken at each base and what the "holes" on each LRU are. Because of the cannibalization assumption, we know that one (or more) specific SRU is missing on every broken unit of the LRU. Similarly we know how many units of the LRU are missing on aircraft at each base.

An outline of the distribution logic for an LRU and its component SRUs is as follows:

1. Find that base, if any, which will be furthest below its allowable aircraft down target, y_j due to this LRU at the end of the planning horizon, even if there is no demand for the LRU during the period. All LRUs in base repair for which there are spare SRUs and all serviceable LRUs in transit to the base are counted as serviceables. Add a unit of the LRU, increasing the base probability, in the following priority order:

   a. Ship the missing SRUs from depot needed to fix a broken LRU at base
   b. Ship a serviceable LRU from depot
   c. Put SRUs on a broken LRU at depot and ship the LRU

Repeat this step until all bases are above y_j (go to step 2) or all assets have been distributed (go to step 4).

2. Same as step 1 except that the asset position is reduced by the expected depot repairable LRU demand during the planning horizon. Repeat until the probability at each base exceeds an input threshold, ρ_0, (go to step 3) or until all assets have been distributed (go to step 4).

3. If the LRU is never base repairable go to step 4. Otherwise compute the expected demand for its first SRU over the planning horizon at each base, and send a unit of the SRU to the base with the smallest probability of meeting its target. Repeat until the probability at each base exceeds the input threshold, ρ_0, or until all serviceable units of the SRU have been distributed. Advance to the next SRU and continue until all SRUs in the LRU are processed.

4. Redistribution - Let ρ_l and ρ_u be lower and upper probabilities specified as input thresholds for redistribution. If there is some base whose probability exceeds ρ_u and another base below ρ_l on a particular LRU or SRU, a redistribution is made from the former to the latter.

After an LRU family is processed, we move to the next LRU family. Note that the distribution algorithm only considers data for one LRU family at a time, and needs no data concerning shops and repair processes. As each LRU family is processed we augment the G's for each base using Equation 8.1 so that at the end we can calculate the availabilities using Equation 8.4.

8.13 Field Test Results for DRIVE

There was a six-month test of DRIVE at Hill Air Force Base in Ogden, Utah during a six month period in 1987. The Air Force estimated that without DRIVE about 40% of the aircraft would have been down after a 30 day period, simulating a wartime effort. If DRIVE had been available for distribution only, about 35% of the aircraft would have been down; with DRIVE repair and distribution about 19% would have been down.

8.14 OVERDRIVE - Separate Distribution and Repair Models

As noted above, I believe that DRIVE would be much simpler with two separate models: one for distribution and a second for repair. This is because the distribution model needs to know nothing about repair shop details, and the repair model doesn't need base availability details. Since the models have to be run at different times, there is no advantage to a single model. I have coined the name OVERDRIVE for this modification.

If we are interested in expected availability targets at each base, instead of the probability of meeting a target number of aircraft down, the procedure for distribution is more complicated than in Section 8.12, because expected availability does not divide into separable functions by item. Instead we need to give each base an essentiality weight. After a trial allocation using the rules above, we need to increase the weights for bases that are below their targets and decrease for bases that are above. We have found that this procedure converges, though it may require 15-20 iterations.

For example, Table 8-3 shows the distribution of F-16 aircraft around the world in about 1987. Suppose that our availability target is 95% at all bases. We invented some data on repaired items, and then allocated them to bases using marginal analysis to determine which base would get the next item.

Note that are large imbalances in Table 8-3, where some bases such as 9, 7, and 8 are under their 95% availability targets by as much as 9%. At the other extreme bases such as 17, 16, and 13 are as much as 4% over the targets. The average error is only 0.38%, but the average absolute error is 2.43%. Now look at Table 8-4, which is the result of lowering the essentialities of bases over their targets and raising the essentialities of bases below the targets. The maximum error is now only 1.88% instead of 9.04%; the minimum error is only slightly changed for reasons discussed below. Lastly, the average absolute error has dropped to 1.45% from 2.43%.

Table 8-3. Availabilities when Bases have Equal Essentialities

Base	# Aircraft	% Avail. Goal	% Avail Actual	Average Error	Average Abs. Error
1	105	95.00	90.87	4.13	4.13
2	18	95.00	88.80	6.20	6.20
3	17	95.00	96.36	-1.36	1.36
4	12	95.00	97.50	-2.50	2.50
5	72	95.00	94.46	0.54	0.54
6	98	95.00	90.43	4.57	4.57
7	70	95.00	88.11	6.89	6.89
8	72	95.00	88.75	6.25	6.25
9	57	95.00	85.96	9.04 Max	9.04
10	125	95.00	91.70	3.30	3.30
11	48	95.00	93.58	1.42	1.42
12	48	95.00	96.01	-1.01	1.01
13	24	95.00	99.51	-4.51	4.51
14	72	95.00	97.04	-2.04	2.04
15	36	95.00	97.95	-2.95	2.95
16	72	95.00	98.38	-3.38	3.38
17	36	95.00	99.58	-4.58 Min	4.58
18	18	95.00	95.29	-0.29	0.29
19	18	95.00	96.00	-1.00	1.00
20	18	95.00	95.14	-0.14	0.14
21	18	95.00	95.82	-0.82	0.82
22	18	95.00	95.53	-0.53	0.53
23	9	95.00	96.06	-1.06	1.06
24	30	95.00	93.67	1.33	1.33
25	18	95.00	95.51	-0.51	0.51
26	18	95.00	95.46	-0.46	0.46
27	17	95.00	97.46	-2.46	2.46
28	18	95.00	96.05	-1.05	1.05
29	24	95.00	95.13	-0.13	0.13
30	18	95.00	95.25	-0.25	0.25
31	18	95.00	95.79	-0.79	0.79
Sum	1242	Averages	93.40	0.38	2.43

Table 8-4. Availabilities when Base Essentialities Change - 16 Iterations

Base	# Aircraft	% Avail. Goal	% Avail Actual	Average Error	Essentiality
1	105	95.00	93.56	1.44	1.89
2	18	95.00	93.51	1.50	4.44
3	17	95.00	93.12	1.88 Max	0.99
4	12	95.00	93.40	1.60	0.70
5	72	95.00	93.70	1.30	0.09
6	98	95.00	93.82	1.18	1.79
7	70	95.00	93.66	1.34	4.32
8	72	95.00	93.81	1.19	1.79
9	57	95.00	93.74	1.26	13.99
10	125	95.00	93.40	1.60	2.70
11	48	95.00	93.70	1.30	0.06
12	48	95.00	95.77	-0.78	0.06
13	24	95.00	99.06	-4.06 Min	0.06
14	72	95.00	94.67	0.33	0.06
15	36	95.00	96.33	-1.36	0.06
16	72	95.00	94.39	0.61	0.06
17	36	95.00	96.83	-1.83	0.06
18	18	95.00	93.26	1.74	1.06
19	18	95.00	93.68	1.34	0.74
20	18	95.00	94.01	0.99	1.10
21	18	95.00	93.75	1.25	0.74
22	18	95.00	92.61	2.39	1.04
23	9	95.00	93.10	1.90	0.53
24	30	95.00	94.18	0.82	1.43
25	18	95.00	93.60	1.40	1.07
26	18	95.00	93.44	1.56	0.81
27	17	95.00	93.95	1.05	0.76
28	18	95.00	93.10	1.90	0.96
29	24	95.00	94.02	1.51	1.00
30	18	95.00	93.49	1.62	1.05
31	18	95.00	93.38	1.45	1.03
Sum	1242	Averages	93.98	0.93	

Average Absolute Error = 1.45

Table 8-4 suggests that we don't have enough stock to satisfy all the bases at 95% availability. Suppose we arbitrarily reduce the availability targets to 90% at the last 15 bases. We use the same iterative procedure and after 20 iterations we obtain Table 8-5. Note that the maximum error is reduced still further to 0.90% and the average absolute error is only 0.81%.[1]

[1] It should be noted that this procedure for adjusting essentialities is not only reasonable, but leads to optimality. The essentialities are Lagrange multipliers of a constrained solution, which reflect the cost of increasing availability at different sites. This is precisely analogous to using site essentialities discussed in the footnote following Figure E-7.

Table 8-5. Availabilities when Base Essentialities Change - Different Targets - 20 Iterations

Base	# Aircraft	% Avail. Goal	% Avail Actual	Average Error	Essentiality
1	105	95.00	95.14	-0.14	2.45
2	18	95.00	95.12	-0.13	8.99
3	17	95.00	94.14	0.86	1.19
4	12	95.00	94.40	0.60	1.03
5	72	95.00	95.01	-0.01	0.12
6	98	95.00	95.02	-0.02	2.72
7	70	95.00	94.92	0.08	4.85
8	72	95.00	95.00	0.01	2.61
9	57	95.00	94.10	0.90 Max	16.00
10	125	95.00	94.98	0.02	4.39
11	48	95.00	96.25	-1.25	0.10
12	48	95.00	96.31	-1.31	0.06
13	24	95.00	99.06	-4.06	0.06
14	72	95.00	95.14	-0.14	0.07
15	36	95.00	96.38	-1.38	0.06
16	72	90.00	93.07	-3.07	0.06
17	36	90.00	95.29	-5.29 Min	0.06
18	18	90.00	89.89	0.11	0.67
19	18	90.00	89.88	0.12	0.73
20	18	90.00	89.71	0.29	0.68
21	18	90.00	89.22	0.78	0.52
22	18	90.00	90.10	-0.10	0.97
23	9	90.00	93.13	-3.13	0.43
24	30	90.00	89.87	0.13	1.26
25	18	90.00	89.99	0.01	0.91
26	18	90.00	89.81	0.19	0.81
27	17	90.00	89.52	0.48	0.60
28	18	90.00	89.91	0.09	0.92
29	24	90.00	89.92	0.09	0.67
30	18	90.00	89.85	0.15	0.64
31	18	90.00	90.13	-0.13	0.93
Sum	1242	Averages	94.05	-0.49	

Average Absolute Error = 0.81

The reason that the minimum errors do not decrease with iterations is that a base with lots of stock and very little demand will have excellent performance even if it receives no stock in distribution. In such cases we need a redistribution capability. We programmed thresholds that could be manipulated by management. No base would receive a redistribution unless its expected backorders were decreased by at least ρ_R and no base would send an item if its expected backorders increased by more than ρ_S. The overall logic is that we attempt to correct existing imbalances before considering the impacts of probabilistic demand. Then we add the demand component, being most concerned about LRUs, since they impact availability directly. After SRUs are considered, we allow for some redistribution to correct long-term imbalances.

It is possible to generalize this procedure to cover SRUs that are common to more than one LRU by considering the "extended" LRU family.

The repair algorithm is much simpler than the distribution algorithm. For each item that uses a given repair shop or test stand, we add across bases to obtain the total expected demand over the planning horizons and the serviceable assets at the end. At each step of the algorithm we find that item with at least one unserviceable unit where the probability of needing an additional unit divided by the repair cost (in man-hours, dollars or whatever) is greatest. The procedure stops when the repair shop capacity is reached.

This repair algorithm can be used to allocate repair dollars or personnel across shops. An ideal allocation of capacity across shops would result in the same ratio of probability/resource for each. Thus the repair algorithm enables us to determine the catch-up requirement across items. Note that the common item problem is handled automatically in the OVERDRIVE repair algorithm logic.

8.15 Current Status of DRIVE

The current status of DRIVE in the Air Force of 2004 is that the model has been implemented at all depots. For the past several years it has been called EXPRESS, and now it is called PARS, Prioritization of Aircraft Recoverable Spares.

PARS still does its prioritization to requisitions. Part of this is due to the configuration concern and part is due to an Air Force policy to do "repair on demand". It was decided that repair on demand meant that the allocations had to be made to existing requisitions.

The prioritization by PARS is really in two parts:

1). Prioritization is done by backorder categories for "high priority" backorders. These would include Joint Chiefs of Staff (JCS) projected coded backorders, Air Force Special Operations Command (AFSOC) backorders and those causing an aircraft to be grounded (Mission Capable Awaiting Parts, known as MICAPs).

2). Remaining requisitions.

The Air Force continues to talk about changing the objective function to availability instead of the probability of no more than a targeted number of aircraft down, but it is still the same as it was.[1]

[1] Private communication from Bob McCormick, formerly a civilian employee at Wright-Patterson AFB and now an LMI staff member.

PARS is used for prioritizing more than 80% of the dollar value of organic repairs at Air Force depots, so it's being used quite extensively. It has reached a level of maturity and acceptance such that it hasn't been necessary to justify its existence lately. The Air Force Material Command (1999) did do a study comparing PARS to a "backorder priority"-based system that the Oklahoma City depot developed. The highlights were[1]:

1). Using a simulation to measure the results over time, the availability-based prioritization scheme (PARS) resulted in about 30% fewer down aircraft under no cannibalization and about 10% fewer down aircraft under full cannibalization

2). Assessing the expected stock levels that would result from using only one day's prioritized list, PARS resulted in about 20% fewer down aircraft under no cannibalization and about 3% fewer down aircraft under full cannibalization.

8.16 Summary

The probability of *y* or fewer aircraft down has been used for many years as the objective function when cannibalization is practiced, because of its analytic simplicity. When there is a single indenture, the logarithms of the cumulative probabilities are convex, and marginal analysis produces optimal solutions.

However, in the two indenture problem the logarithms of the probability of *y* or fewer aircraft down are not even additive separable functions of the items. Thus, they are not convex, and there is no guarantee that marginal analysis will produce an optimal solution. More importantly mere are several conceptual drawbacks to the use of mission accomplishment as the objective function, even for a single indenture.

It is possible to use expected availability as the objective function when cannibalization is practiced. There is an elegant procedure that enables us to use marginal analysis, even though the objective function is not separable into independent item calculations. Though there is no guarantee that marginal analysis produces optimal solutions, there is empirical evidence that the procedure leads to near optimal policies. Furthermore, for a given budget the policies produce higher availabilities than under the probability of *y* or fewer aircraft down.

For a specified budget the optimal stockage policy under the availability objective function tends to look similar to that under the probability of *y* or

[1] Private communication from Richard Moore, a civilian employee at Wright-Patterson AFB.

fewer aircraft down for some value of y. As the budget is increased, the policy looks similar to that under the probability of y or fewer aircraft down for a smaller value of y. This seems intuitively reasonable. The advantage of the expected availability objective is that the user need not guess the value of y that is appropriate for the budget available. On the other hand, we would recommend solutions for the probability of y or fewer aircraft down at several values of y, evaluating the availability and the expected LRU backorders for each. The computation of the probability of y or fewer aircraft down is much simpler and we can look at the tradeoff between availability and a proxy for cannibalization workload.

We did not discuss any modeling of redundancy in this chapter. However, in the simple case where there is only one end-item such as a nuclear power plant, the theory is easily adapted to cover the case where the plant must shut down unless at least K of the N copies of a system such as water circulating are operating (see Problem 6).

In this chapter we have described a model for depot repair and distribution known as DRIVE. This model differs fundamentally from a procurement model in that decisions are based not on steady-state probabilities but on probabilities computed over a planning horizon. Regardless of the procurement model used, DRIVE has the capability to make better short-term decisions because they are based on the actual location and condition of items rather than steady-state probabilities.

8.17 Problems

1. Show that Equation 8.2 is convex for Poisson, negative binomial or binomial probabilities. *Hint*: Note that for the Poisson of Equation 2.4

$$\log[1 + p(x)/P(x - 1)] > \log[1 + p(x + 1)/P(x)]$$

if and only if

$$[p(x)/P(x - 1)] > [p(x + 1)/P(x)]$$

and show that the latter is true because after cross-multiplication, each term on the right-hand side is dominated by a term on the left. For the negative binomial of Equation 4.5, the same procedure is followed, after an easy proof for $a > 1$ that

$$\binom{a+x-1}{x}\binom{a+x-y}{x-y+1} \geq \binom{a+x}{x+1}\binom{a+x-y+1}{x-y}$$

for y in the interval $1 \le y < a$ and any x.

For the binomial we need only show that for any y where $n \ge x \ge y$:

$$\binom{n}{x}\binom{n}{x-y} \ge \binom{n}{x+1}\binom{n}{x-y-1}$$

2. Calculate cumulative Poisson probabilities for means of 1 and 10 and then use Equations 8.1 and 8.4 to verify that the stockage policy optimized for cannibalization in the rightmost column of Table 1.4 yields an availability of 96.04% when cannibalization is practiced.

3. By substituting Equation 8.5 into Equation 8.6, show that any trial solution with stock levels $s_0, s_1, \ldots s_J$ such that $s_1 \ge Z_1$, $s_2 \ge Z_2, \ldots$ and $s_J \ge Z_J$ can be improved by building an LRU from the SRUs times quantity per LRU (Z's). This assumes that the J SRUs times their quantity in the LRU comprise the entire LRU, and that their total cost is the same (or greater than) that of the LRU.

By repeated application of this result, it is clear that a final solution is achieved only when at least one of the SRU stock levels falls below its quantity in the LRU (QPA). Note that the result applies to both objective functions and any probability distribution of demand. This is a useful fact that simplifies the search for optimal solutions.

By contrast it is easy to show that a similar result does not occur in the non-cannibalization case. The reason is that in that case an LRU can be backordered because of a single SRU.

4. Solve problem 10 of Chapter 2 when cannibalization is allowed, and contrast the solution with the no-cannibalization solution from Chapter 2.

5. Solve problem 11 of Chapter 2 when cannibalization is allowed, and contrast the solution with the no-cannibalization solution from Chapter 2.

6. Problem 6 of Chapter 5 addressed the problem of modeling redundancy when there is a single end item as in a nuclear power plant. Consider the same problem when cannibalization is allowed, and show it can be accommodated with minor changes to Equation 8.1.

7. Show that the mean and variance for the number of aircraft down due to noncannibalizable first-indenture items are approximately equal to the sum of their backorder means and variances respectively. *Hint*: Use Equation 2.19 and the Taylor series expansion for the exponential.

8. (Research) Use simulation in order to develop an estimating relationship between expected LRU backorders, as discussed in Section 8.5 and Table 8-2, and cannibalization workload. The relationship must depend on variables such as the number of aircraft (since the number of possible cannibalizations increases with fleet size) as well as other variables to be identified in the research.

Chapter 9

APPLICATIONS

Any sufficiently advanced technology is virtually indistinguishable from magic.

-Arthur C. Clarke

9.1 Chapter Overview

In this chapter we describe a number of different applications of the VARI-METRIC theory. Even though the basic theory assumes that resupply or repair can be initiated whenever a demand occurs, it turns out that the theory is appropriate for a much wider set of problems.

The most important application is to *commercial airlines* (Section 9.2), which depart from the usual assumptions of VARI-METRIC in that availability is a meaningless objective for aircraft that move around a set of airports. A related problem in Section 9.3, which is of particular importance to airlines, is *redistribution of the spares already owned* before putting in a new procurement, including the possible sale of items that are not needed.

Then in Sections 9.4 to 9.7 we turn to variations from VARI-METRIC in the repair and resupply assumptions: *periodic resupply, no resupply, repair-in-place,* and *contractor repair.* Even though we went into a lot of detail in Chapters 6 and 7 to develop the theory for periodic resupply of the space station, much of that was because of our desire to model redundancy; most periodic applications are easily handled by the VARI-METRIC theory. In Section 9.8 we estimate the *probability distribution for repair delay.*

In Section 9.9 we consider changes to the multi-echelon structure; namely where a given *site may be an operating site and also a support site* for other sites. This is followed by situations affecting the indentured parts hierarchy. In Section 9.10 we consider situations where the *indentured structure is unknown* or should be suppressed in the interests of simplified data management. Section 9.11 shows a case where *two model runs are required to solve a complex system.* The problems with items that have *limited interchangeability and substitutability* are addressed in Section 9.12. We show in Section 9.13 how a *limited redundancy capability* can be used to model both cannibalization and no cannibalization. In Section 9.14 we deal with the case where unfilled demand may not generate a backorder.

9.2 Airline Applications

The biggest obstacle in using VARI-METRIC theory for airline operations is that the availability objective is meaningless. We cannot compute the percent of available aircraft at each location, because aircraft are not assigned to particular locations; they move around.

A figure of merit used by airlines is the on-time departure rate (for spares), which we can calculate as follows:

$$\text{On-Time Departure Rate} = \frac{\text{Departing Flights - (LRU Non-Fills)(Crit.)}}{\text{Departing Flights}}$$

where departing flights are those scheduled and where we can make the simplifying assumption that any LRU demand that can not be filled results in an aircraft not departing on time; this last assumption is reasonable because the demand for spares in a commercial airline operation is an order of magnitude less than for military aircraft, and the probability of multiple failures on one aircraft is very small. We allow the number of LRU non-fills to be multiplied by a criticality number; this is usually one but might be less for certain items that are not required for every flight, e.g. equipment needed for flights over water.

Thus the objective function is basically the first-indenture item fill rate at operating locations, though a 70% fill rate might translate to a 90% on-time departure rate.

A version of the VMetric model, described in Appendix E is been used by several airlines. Unfortunately, we cannot identify the airlines or precise savings for competitive reasons, but a typical result over nine commercial airline fleets has been a *40% reduction in spares cost and a 20% improvement* in performance.

One airline went from an on-time arrival rate of 58.6% and an average delay of 53 minutes to an 86.6% on-time arrival rate slightly more than a year later, attributing this industry leading performance to "new automation to optimize the provisioning of spare parts".

9.3 Redistribution and Sale of Assets

One of the first problems facing the airlines, as well as other users, is how to redistribute owned stock by item optimally across locations, allowing purchase and sale of stock up to any specified target. The ability to sell spares that are not needed is particularly important for the airlines where the resale price of a spare is often a significant fraction of its original cost.

It turns out that the optimal solution is quite straight forward. The model begins with no stock and uses the sales price of each item in the optimization. When it exhausts the number of units owned on any item, it converts to the purchase price of that item for subsequent iterations. When the optimization target (budget, on-time departure rate, or whatever) is reached, the net stock on any item whose owned stock has not been exhausted is marked for sale. It is easy to see that this procedure is optimal. The spares for sale were not attractive to the model even at their resale prices.

There is one other detail for items with a positive condemnation rate. We do not want to sell all units of these items not needed by the model; instead we perform a calculation of how much to reserve for future condemnations.

9.4 Periodic Resupply

The VARI-METRIC theory of Chapter 5 can be used to model periodic resupply as well as continuous resupply. We did present theory in Chapter 6 for the periodic resupply of the space station, but this more elaborate theory is not required unless the spares optimization can plan on having certain "holes" in LRUs that do not cause end-item failure because of redundancy.

The important thing is to properly interpret the VARI-METRIC inputs for the periodic resupply case. For example, suppose we have a fleet of 50 submarines which are visited on average every 30 days by a supply ship. We assume that the submarines can do only remove-and–replace maintenance. Let's assume that everything can be fixed on the supply ship in an average of 45 days.

This is a two-echelon problem with 50 operating sites. We need to specify the Not-Repairable-This-Station (NRTS) for each site, which is one minus the repair rate. The NRTS = 1 for each submarine and item; NRTS = 0 for the supply ship. The repair time for the supply ship is 45 days. Before looking below, try to estimate what the order-and-ship time should be. Remember that the order-and-ship time is the average time from the failure of an item until its replenishment by our supply source *when the latter has a spare on the shelf.*

It turns out that there are two possible answers. As in the space station, the availability, the probability of a submarine being fully operational, decreases over time. If we want the average availability, we should use 15 days as the average time between needing a spare and being resupplied. On the other hand, if we want the minimum availability that will be reached just prior to the arrival of the supply ship, we should use 30 days.

Let's consider a slight variant of the problem above, where our country can only afford one submarine. What changes, other than the number of end-items?

Again this is like the space station case where repairs were assumed to take 160 days, but if the shuttles had a 180 day periodicity, the repair time was effectively 180 days. Similarly, the supply ship repair time is effectively 60 days here for the same reason.

The case with one submarine brings up one other point. Where will the model put all the spares? Unless we constrain the stock held on the submarine by using maximum stock levels, or put a "shadow price" on volume, the model will put all the spares on the submarine, because that is where all the demand occurs.

There is another application which can be treated the same way. The Israeli Air Force has radars at remote sites that are periodically visited by maintenance teams. During these periodic inspections, any failed items are replaced, assuming a spare is available. We assume that because of redundancy the radars are not usually down even though there are some "holes" (it is likely that there would be some monitoring devices on the radars to indicate a failure in the system).

9.5 No Resupply: Flyaway Kits

There are important applications where there is no resupply. The military services determine the spares needed for flyaway kits or war reserve spares kits that must support independent operations for a period of time. We show below that the VARI-METRIC theory developed above is appropriate for the no resupply case, provided that cannibalization is not practiced.

The no resupply case implies a single site. Consider a first-indenture item that is not repairable. The expected backorders are given by Equation 2.7 as always, but the random variable due in, DI, is replaced by the random variable X, the number of demands during the period. If demand is Poisson the probabilities are Poisson with mean $m\tau$, where τ is the length of the period, instead of mT, where T was the mean repair time. Because there is no repair, Palm's theorem is not needed. As in the previous section for periodic resupply, the mean $m\tau$ would give the minimum availability, reached at the end of the period, whereas $m\tau/2$ would give the average availability.

More generally suppose that the item has a probability r of being repaired. Assuming that the supply of second-indenture items is adequate, the expected backorders for the first-indenture item are given by Equation 2.7 with a mean that is equal to $m(rT + [1 - r]\tau)$.[1]

Second-indenture items make an impact through the value of T as in VARI-METRIC. In principle the only complication is that the times used in VARI-METRIC must be an appropriate combination of the repair time and the length of the period, reflecting the probability of repair for each item. The typical military application usually assumes cannibalization, and this was addressed in Chapter 8.

9.6 Items that are Sometimes Repaired-in-Place

Sometimes items are repaired-in-place, reducing the requirement for spares. For example, suppose the average annual demand for an LRU and its two SRUs are as shown in the left half of Table 9-1, where the indentations indicate second-indenture items (note that the demand for the two SRUs may exceed the demand for the LRU). Let's assume that the LRU can be repaired-in-place 30% of the time. What changes are required to reflect this capability?

Table 9-1. Illustration of a.) Normal Demand; b.) 30% LRU Repair-in-Place

Normal Demand		30% Repair-in-Place	
Item	Annual Demand	Item	Annual Demand
LRU	10	LRU	7
SRU #1	8	SRU #1	5.6
SRU #2	4	SRU #2	2.8

[1] In this expression we assume that $\tau > T$. In order for this expression to be correct for any τ, T should be replaced by $\min(\tau, T)$. Note that when $r = 0$, indicating no repair, the mean reduces to $m\tau$ as in the previous paragraph.

	SRU #1	2.4
	SRU #2	1.2

Clearly the demand for the LRU is only 70% of its original value, and those demands break down into SRU demands as before. But, we have demands for SRUs themselves which add up to the original totals. Each SRU will now appear twice as a common item in the indentured list of items, once as an SRU and once as an LRU.

9.7 Contractor Repair

Let's assume we are operating a fleet of aircraft, and that all repair is done by a contractor at its own site. The spares at the contractor site are calculated and owned by him. He contracts with us to repair and return items to us in an average of 60 days. How do we feed VARI-METRIC so that it will determine the spares we need at the operating sites to meet some availability target?

This is a one-echelon problem since the repair site is not part of our organization; thus there is no order-and-ship time since that pertains to times within a multi-echelon system. The 60 days looks like a procurement lead time, where we are assuming the 60 days includes the time to ship the broken item back to the contractor. The Not-Repairable-This-Station (NRTS) for each operating site, which is one minus the repair rate, must be NRTS = 1. Of course, we could generalize this to NRTS < 1, where some repair is performed at the operating sites.

How does the contractor determine his spares? He runs a one-echelon calculation, too. The 60 day average repair and return guarantee by the contractor implies that the average delay per demand plus the time, t, to get the item to the operator should be 60 days. To be on the safe side, we may want to include the retrograde time for the broken item as well as the shipping time in t. Then if we are using a model such as VMetric, which allows this objective function, we set a target for the average delay/demand for LRUs at operating sites of $(60 - t)$.

9.8 Probability Distribution of Delay Time

One of the VMetric targets, as noted in the previous section, is the average delay per demand for LRUs at operating sites. Suppose we are interested in the probability distribution of delay on individual LRUs at any site. The probability of no delay at a site is just the item fill rate, which is one of the outputs in VMetric.

Of course, if we are a contractor making guarantees about our support we may want some information about the probability distribution of repair time, not just the average delay and not just the probability of no delay.

It turns out that this can be done, but it is necessary to assume something about the shape of the repair distribution (Palm's theorem does not apply). Results are easy to compute for Poisson demand in two cases: exponential repair times (Higa et al. 1975) and constant repair times (Sherbrooke 1975).

For the constant repair time case, the delay, t, can never exceed T, since T is assumed to be constant. Now consider delays of t or less where $t \le T$. If s - 1 or fewer units were demanded during the previous time interval of length $(T - t)$, the delay can not exceed t, regardless of how many demands were placed earlier. Thus, the formula looks very similar to that for the expected fill rate in Equation 2.6 except that the mean is not mT, but $m(T - t) \equiv x$

$$\Pr\{\text{Delay} \le t\} = \sum_{i=0}^{s-1} e^{-x} x^i / i! \qquad 0 \le t \le T$$

and the delay is zero otherwise. The general formula for the average delay is given in Higa et al. (1975); in the constant repair time case it is most easily evaluated by numerical integration. They provide specific formulas for the $\Pr\{\text{Delay} \le t\}$ and the average delay in the exponential repair case.

In Table 9.2 we compare the results for exponential and constant repair times where t = 10 days (0.027 years) and the average annual demand is 3.65. We note that the probability of delay of 10 days or less is quite similar for both repair distributions, although the repair distributions themselves are extremely different. The average days of delay are very similar as well. The probability of no delay, the fill rate, is identical because of Palm's theorem.

Table 9-2. Probabilities of Delay under Exponential and Constant Repair Times

			Exponential		Constant	
Stock Level	Average Repair Days	Prob. No Delay	Prob. Delay≤10 Days	Average Days Delay	Prob. Delay≤10 Days	Average Days Delay
1	5	0.951	0.999	0.125	1.000	0.123
	10	0.905	0.986	0.494	1.000	0.484
	20	0.819	0.928	1.940	0.905	1.873
	40	0.670	0.783	7.508	0.741	7.031
2	5	0.999	1.000	0.002	1.000	0.002
	10	0.995	1.000	0.016	1.000	0.016
	20	0.982	0.996	0.123	0.995	0.121
	40	0.938	0.968	0.908	0.963	0.877
3	5	1.000	1.000	0.000	1.000	0.000

10	1.000	1.000	0.000	1.000	0.000
20	0.999	1.000	0.006	1.000	0.006
40	0.992	0.997	0.086	0.996	0.084

This suggests that the constant repair time solutions may be good enough for most purposes; if more accuracy is desired for an arbitrary repair distribution, we would recommend averaging the two results.

9.9 Sites that are Both Operating and Support

There is a large airline which operates aircraft from Washington National (Reagan), Washington Dulles, and Baltimore-Washington, all in close proximity. Washington National is also a support site for all three operating bases. It is easy to represent this situation in the required tree-structure. We break Washington National into two sites: an operating site and a support site. We set a maximum stock = 0 for all items at National operations so that the stock for National support is available to all three operating sites (there is only one stockroom at National). The NRTS = 1 and the order-and-ship time = 0 (or something very small) for all items at National operations.

This situation is easy to handle, but it is automated in VMetric.

9.10 Large Systems where Indenture Information may be Lacking

There are applications where the parts hierarchy information is unknown. An example is the Defense Logistics Agency (DLA) which provides common spares to all military services. DLA has knowledge about the system or systems that use an item, and it may be able to classify items into different criticality groupings. But it does not usually have the parts hierarchy data to properly classify an item by indenture, and utilize the detail of VARI-METRIC theory. Furthermore, it would not be possible to compute a meaningful availability, because they have responsibility for only a fraction of the items on a given system. Instead the managers rely on various fill rate targets by item criticality class.

In our view the VARI-METRIC theory should still be applied to items grouped by criticality class. All items would be considered first-indenture, since no other information is available. The model theory uses backorders in the determination of optimal stock levels, but we can calculate the fill rate at each point on the cost-backorder optimal curve. Thus we can calculate the item stock levels in a multi-echelon setting that will produce a target fill rate at the base level. In this way we still get the benefits of the VARI-METRIC theory which provides greater safety level on lower cost items and for the retail, base locations in a systematic way.

Even when indenture data is available, it does not make sense to put all items on a system into one massive computation. The lower indenture bits and pieces will have extensive commonality, and the data input will become very complicated. This is particularly true if there are many indentures, because at the lowest indentures it is possible that a given common item appears many times in the indenture hierarchy after its several "parents".

We believe that the multi-indenture version of the theory should be run for the major items whose cost is appreciable. But the bits and pieces should be run in a separate one-indenture computation to some target such as fill rate. Another useful objective function that we have used instead of fill rate, or in combination, is the average delay per demand; this is the backorders divided by the demand, of course, but it puts the backorder level into a more meaningful context.

The problems with DLA managed items are discussed further in Section 10.6 concerning model implementation by the Air Force.

9.11 Systems Composed of Multiple Sub-Systems

Suppose that we have a system with 1000 hand-held receivers in the field. The receiver MTBF is 180 days. After a receiver breaks, it takes an average of 2 days for the user to bring it in for service to one of 10 repair locations. The user will be given a new receiver if one is available. Sometimes the repair location can fix the receiver and sometimes it has to be shipped to a depot.

The system also includes a transmitter at a central location which is to be spared as well. It is desired to have a 95% availability for the overall system.

Consider how this might be modeled. First, note that even with infinite spares, the availability due to the receivers can never be 100%. This should make us think of maintenance availability = 100 x MTBF/(MTBF+MTTR) = 100x180/182 = 98.9% where MTTR is the mean time to repair. This implies that to get an operational availability of 95%, the supply availability must be = operational availability/maintenance availability = 95%/98.9% = 96.1% (see discussion following Equation 2.17).

Unfortunately, we cannot just make an optimal VARI-METRIC run on all the spares to a target of 96.1%, because there are 100 receivers and only one transmitter. We need to make two VARI-METRIC runs, one for the receivers and one for the transmitter, where the product of the two availabilities is about 96.1%. The selected solutions from each of the two curves should also have the property that the slopes of the individual availability-cost curves have about the same slope, indicating that the marginal analysis values at the solution are similar.

9.12 Items with Limited Interchangeability and Substitutability

In the real world there are items which are members of a group with limited interchangeability and substitutability. Suppose there are two LRUs, either of which can be used on the aircraft, or perhaps LRU #1 can be used on a certain group of aircraft and LRU #2 can be used on another group, perhaps with some overlap. It may be that one of the LRUs is a newer version which is slowly replacing the older LRU.

Let's assume that LRU # 1 has 70% of the demand, and LRU #2 has 30%. The correct way to represent this is in VARI-METRIC is to put both LRUs into the indentured parts list, with LRU #1 getting a demand rate of 70% of the total and LRU #2 getting 30%. Of course, each LRU may have a different internal structure with different component parts.

In the absence of configuration control information, we are forced to assume that each site has the same 70/30 mix of the LRU. But, if we do know which type LRU is used on each aircraft, then we can input site specific demand rates for the two types of the LRU.

9.13 Redundancy

We built a limited capability to model redundancy into the VMetric implementation (described in Appendix E) of VARI-METRIC. The major limitation is that there can be only one end-item. As an example, let's consider a nuclear power plant with four water circulation pumps, of which at least two must be operating for safety reasons. (This is actually the problem that we faced when we decided to include some redundancy modeling). When there is only one end-item, the modeling is quite straightforward because if we only need 2 of 4 copies, say, of the LRU to be operating, it is like having two free spares.

Let's suppose that each pump consist of two LRUs. We consider two cases: (1) cannibalization of the LRUs is allowed (or some kind of cross strapping as in the space station electrical system); (2) no cannibalization.

In the first case of cannibalization all we need to do is list the two LRUs as first-indenture items with $K = 2$ and $N = 4$ where K is the minimum number that must operate and N is the quantity of the LRU on the end-item (power station). We don't need to separately identify the fact that these two LRUs comprise a pump.

In the second case of no cannibalization we need to introduce a *dummy item* at the first-indenture and demote the two LRUs to SRU components of the dummy, which we will call "Pump". The dummy item will have $K = 2$ and $N = 4$, whereas $K = N = 1$ for each of the components. We can think of the two components as connected together in pairs with no possibility of cannibalization; if either one has a "hole", the pump is not operational.

Of course, we do not want to buy any units of the dummy item, so its maximum stock should be set to zero. All demand for the dummy should be passed to its children, so the NRTS = 0 and the site repair time = 0. As one might expect, the first case of cannibalization is typically much less expensive. On the other hand, most logisticians are unwilling to allow some "holes", even when there is redundancy, when they are planning what spares to purchase.

9.14 Unsatisfied Demand may not be a Backorder

Consider a commercial airline that wants to use VARI-METRIC theory for optimizing its sparing. But when a demand cannot be filled or obtained from the contractor source, the airline has the option of buying the item on the open market. This has some of the characteristics of the lost-sales case, mentioned earlier, because a backorder is not established. How should this be modeled?

This is not the classic lost-sales case, as discussed in Hadley and Whitin (1963), in which there are never backorders; sometimes there will be lost-sales and sometimes there will be backorders. It would be very difficult to build a VARI-METRIC type model to exactly model this mixed situation. Even the classic lost-sales case of Hadley and Whitin is substantially more complicated than the backorder case.

In our view the VARI-METRIC approach is appropriate in this application. It is true that VARI-METRIC will tend to buy somewhat more stock than would be optimal if the lost sales that sometimes occur could be easily taken into account. However, our objective is to keep those lost sales low, and the advantages of multi-echelon, multi-indenture models are substantial.

9.15 Summary

We have presented a number of applications where the basic VARI-METRIC assumptions have been modified, but without affecting the applicability of the theory.

Chapter 10

IMPLEMENTATION ISSUES

Our little systems have their day.

-Tennyson

10.1 Chapter Overview

In this chapter we consider some of the issues relating to implementation of stockage models. We first discuss comparisons of VARI-METRIC with existing stockage policies, and point out errors favoring VARI-METRIC that are sometimes made.

Then we address the amount of data required by the models, and the difficulty in keeping the information current. We distinguish between two types of data: (1) primary data such as demand rates that relate to the "physics" of the problem; (2) derived data such as average repair times that are influenced by management decision-making. We argue that standard times may be more appropriate than measured values for derived data.

We discuss the effect of errors in the data on the model decisions. In many cases the estimation process is robust in the sense that large errors in the data result in small errors in the decisions. This enables us to concentrate on those data elements such as demand rates and variances whose impact on stockage decisions is most critical.

We stress that the theory developed in this book should be used creatively in the assessment of alternative support policies. It should not be assumed that the logistics system is fixed, and that the logistician's job is to optimize the spares only for that configuration. Are there ways to reduce order-and-ship times, lead times, and repair times and are they cost-effective? Is it cost-effective to increase the percentage of items repaired at a base by means of more capable test equipment? Should there be a two echelon structure, or does it make sense to have a third, intermediate echelon between the bases and depot? Perhaps a ragged echelon structure as depicted in Figure 1.3 is appropriate.

The military services of the United States and other users have implemented the theory in this book in different ways for different reasons. Hopefully our brief review will make the reader realize that there are many options in how these techniques can be utilized.

Although it may be possible to include every item on an aircraft in a massive multi-echelon, multi-indenture computation, the problems of maintaining the data base and all of the commonality information are formidable. We argue that it makes sense to run the multi-echelon, multi-indenture model to an availability target for the group of important items; that a single-indenture run be made to a fill rate target for the group of less-important items that would otherwise require extensive data input to represent commonality relationships.

We discuss model hierarchies, noting that early in the planning for a system, models will be crude and may even be deterministic. As the design is finalized, it is appropriate to consider more detailed models, such as those in this book for spares support. During the operations phase, our models can be still more detailed taking account of the latest information on location and condition of spares. For wear-out items, the knowledge of the age of each installed component can enable us to make even better tactical decisions of what to repair and where to distribute material.

Finally, we summarize by noting the number of different ways in which the system approach has been used in this book. In fact, that is what this book is really about. The system approach is more complicated than the item approach. It requires the manager to review the cost-availability tradeoff curve, and select an appropriate target. It requires the analyst to determine stock levels for each item and location that take into account the stock levels for other items and locations, as well as the system target. In statistical estimation, the system approach tells us that we should not limit ourselves to data on a given item when we are making estimates of its mean and variance.

But the system approach is not just a different perspective than the item approach. It leads to different stockage decisions and significant improvements in system availability and cost.

10.2 Comparison of VARI-METRIC with Other Stockage Policies

This is a critically important section. It is easy to fall into a trap when comparing VARI-METRIC with existing policies, in a way which inflates the advantages of VARI-METRIC. We did make some valid comparisons in the George AFB field test of Section 1.9 and in our tests of various demand prediction techniques in Appendix C. Recall that in those cases we used data from a base period, made predictions, and then evaluated those predictions with new data.

Another common approach is to take existing stock levels and compare them with levels computed by VARI-METRIC, usually assuming constant Poisson demand. There are a couple of problems: (1) the existing stock levels were computed from earlier data, and if we believe demand rates "drift" over time, then we have had the advantage of more recent, and relevant data; (2) we know demand rates do not stay constant, so the assumption of constant Poisson demand is almost surely wrong. By assuming constant Poisson demand and *known, error-free demand rates,* we automatically give VARI-METRIC an advantage; it is possible that the existing stock levels would perform better if "real data" had been used to assess their performance. Not only is this procedure wrong, but it tends to overstate availabilities or understate the required budgets.

In addition, it is useful to look at the existing policy to identify elements that might lead to non-optimality. It is not really sufficient just to show that VARI-METRIC is better; users should want to know why, if possible. For example, in the Air Force field test of Section 1.9 we noted that the Air Force policy didn't even consider unit cost. While there were other reasons for the Base Stockage Model success, including the "objective Bayes" procedure for demand prediction, it is not surprising that our policy was more cost-effective than a policy that ignored unit cost. More generally, stock levels should not be insensitive to echelon and indenture differences.

10.3 Use of Standards versus Measured Quantities

The models in this book share one characteristic. They tend to have a voracious appetite for data. The importance of the data quality to decision-making varies, though demand rate estimates are certainly the most crucial.

Data quality tends to improve with time. Early in program planning, before any system has been fielded, the data are usually very "soft". But we have argued above, that for purposes of overall budgeting, the cost-availability curves from this initial data should be meaningful unless there is a systematic bias in the estimating procedures.

As experience data begin to accumulate, it seems only natural that initial estimates should be revised. For example, in the case of demand rates Bayesian analysis provides a natural mechanism to shift from initial estimates to real data. Of course, in demand rate estimation it is important to use different procedures for items whose primary failure mode is thought to be due to wear out. For wear-out items, we may want to increase demand rates if early failures exceed expectations, but we may not want to decrease demand rates if early failures are low.

However, there are some data elements that should not necessarily be revised to agree with measured values. Consider the base repair time for an item that is base repairable. Suppose that the standard used in the model initially is four days (assuming that repair parts are on hand), representing engineering judgment and some allowance for queueing delays. But suppose that the measured average is ten days. Obviously if the model is rerun for a repair time of ten days on this item, a larger fraction of the total budget will be allocated to this item, other things being equal.

It is important to know why the measured time is ten days. If it is because the standard time was incorrectly estimated originally, then it makes sense to modify the model data. However, if we bought more than enough spares for this item and the item has a ten day average repair time because we don't need to fix it in any hurry, it would be inappropriate to change the repair time in the model. By increasing the repair time in the model, we would buy even more spares for an item whose supply is already adequate.

The key consideration is to divide our data into two categories: (1) primary data such as demand rates and unit costs that are not influenced by management decisions; (2) derived data such as average repair times that depend on management decisions. In the former category, initial estimates should be revised as experience accumulates; in the latter the revision of initial data must be made only with a full understanding of the management policies. In many cases it will be more appropriate to use standards for derived data values.

10.4 Robust Estimation

When a statistician talks about *robust estimation methods,* it is a reference to techniques where large changes in the data result in small changes in the decision variables, e.g. stock levels. We have seen a number of instances of robust estimation in the book. For example, in the economic order quantity of Equation 1.1 an error in the demand rate or item cost by a factor of 4 is reduced by the square root to an error of 2 in the order quantity. In Table 6.5 we saw that with 8 identical copies of a system and a fixed budget, the optimal set of stock levels corresponding to a requirement that at least K of the eight operate was virtually unchanged for values of K between four and eight.

From Table 1.4 we can infer that an error in the estimated cost of an item by a multiple of 10 translates to an error in stockage by a multiple of something over 2 (assuming Poisson demand and no cannibalization in the optimization). The error is less if the optimization assumes cannibalization; more if the variance-to-mean ratio exceeds one.

As we have said many times, the estimation of demand rates and variances is the most critical. While logisticians typically understand the importance of demand rate estimation, they tend to be less aware of the importance of variance estimation. In Chapter 4 (Table 4.4) we compared the results of assuming a variance-to-mean ratio of one (Poisson) with a variance-to-mean ratio that increases with the mean according to a power curve relation, and showed that the latter did substantially better over the next year for the C-5 aircraft. Similar results were noted in Appendix C for other aircraft types.

However, we have heard logistics planners say that the power curve relation cannot be used, because the required budget is too large. Our view is that we should make the best estimate of every parameter required in the models, including the variance-to-mean ratio. If our budget in the case of Table 4.4 is $80 million (or less), it still makes sense to spend that money in accordance with the best estimates of the "physics" of our problem. If we spend the money on the optimistic assumption that we know the demand rates precisely and they will not change in the future, we should not be surprised to see severe degradations in performance, as in Table 4.4.

10.5 Assessment of Alternative Support Policies

It is the author's hope that the modeling techniques described in this book will not be used passively by logisticians. That is, it is possible to assume that the logistics support structure is fixed and that the task is to allocate a specified budget optimally, or achieve a specified availability at minimum cost.

A far more creative and interesting role for the logistician is to identify and test interesting support alternatives. For example, are there ways of reducing order-and-ship time, and what would the overall cost and benefit be? The cost might be due to more use of expedited (and presumably more expensive) transportation, and the benefit can be assessed by running the model with a different order-and-ship time.

Typically the answer is not simply "yes" or "no". It usually depends on the target availability level. Other things being equal, the benefit of shorter order-and-ship times (and similar alternatives) is greater if the target availability is high. We saw an example of this in Figure 7.5, where the benefit of commonality on the solar array increased with the availability target. In Problem 1of Chapter 10, the reader is asked to speculate on how various factors affect the decision to put additional test equipment at the bases.

More generally, there are logistics implications of different operational policies. We are not suggesting that logistics should drive operations, but there should be some feedback from logistics to operations. In the 1960s we were asked to estimate the war reserve spares required so that a certain number of massed sorties of six aircraft could be mounted from a total fleet of six. As we were assured that six aircraft was a meaningful number for a massed sortie, we pointed out that if it were possible to have a total fleet of seven aircraft from which six aircraft would be chosen, the spares requirement would be dramatically less (and different). The point is that operational planners should be aware of logistics implications and alternatives.

Last, we want to reiterate a point made earlier concerning the judgment of logisticians. There may be situations where the logistician is aware of factors that are not considered by the model. Perhaps there is a good reason to have a minimum (or maximum) stock level on an item. If so, these constraints should be included in the model optimization. The model results should provide useful information, but they should not prevent the logistician from applying expert knowledge as appropriate. Of course, this means that the logistician needs to understand what is considered by the model and what isn't. That is the reason for this book.

10.6 Model Implementation – Air Force

The military services in the United States have implemented the theories in this book in various ways. The Logistics Management Institute programmed a version of VARI-METRIC known as the Aircraft Availability Model (AAM) and described by O'Malley (1983) that has been used to estimate spares budgets for headquarters personnel in Washington, D.C. It has been programmed for follow-on provisioning by the Air Force Logistics Command in Dayton, Ohio. However, they assume that all bases are identical for procurement purposes.

The Air Force Logistics Command Regulation 57-4 that deals with Recoverable Item Management specifies that MOD-METRIC is one of the acceptable procedures for initial provisioning. The Air Force has used it for some procurements of new engines, and to determine an optimal mix between the number of whole engines and first-indenture engine items. However, in most cases the Air Force does not use an optimal availability model in initial procurement. Instead they buy only enough stock to fill the repair pipelines and some stock to cover condemnations for a short period of time. They buy little, if any, safety stock. They depend on extensive contractor support, and have low target availability goals at the start of a new program. They hope to get better estimates of demand and other parameters before having to make their first real follow-on buy using an optimal availability model. Any items with extremely high initial failure rates are candidates for redesign.

In principle the follow-on procurement should use as input the assets of every item, and optimize the additional procurement necessary to meet availability targets. As a practical matter this is difficult because of the way in which data is processed. Typically the gross requirement is computed from the model, and at a later time the assets in the system are subtracted to obtain the net requirement.

The Air Force in 2004 is finding that over half of the aircraft downtime is due to consumables; in the past it was tacitly assumed that repairables were the biggest cause. Moreover, numerous studies show that roughly 50% of all aircraft down hours (MICAPs) are directly caused by Defense Logistics Agency (DLA) managed parts. That doesn't count the additional MICAP hours for Air Force parts being driven by shortages in DLA repair parts. The problem is, of course, that DLA does not have the indentured parts information to do an optimal indentured spares computation. Another problem is that DLA owns the wholesale stocks, and the Air Force owns the retail. Since it does make sense to improve the stock leveling for DLA managed consumables, the Air Force is experimenting with procedures for computing and transmitting these levels to DLA. Recently the Air Force has been able to reduce actual backorders by 65% at the depot retail supply accounts with no increase in cost.

Of course, consumables can be handled in VARI-METRIC, but as we suggested in Section 9.10, even when the indenture information exists, it does not make sense to put every screw and washer, and its indentured relationship to many different parents into the optimization. Important consumables in terms of indenture and demand rates should go in the indentured computations, but a single-indenture computation for all other consumables is recommended.

We remind the reader who is interested in the Air Force applications that DRIVE (now PARS) for depot repair scheduling and distribution was discussed at length in Sections 8.8-8.15.

We should also mention a study performed on the C-5 aircraft (TFD Solutions, 2000).The study considered 742 repairable components that had demands both from the field and the overhaul facility. The VARI-METRIC theory was used to evaluate the existing aircraft locations and the components placed where the USAF RBL (Readiness Based Leveling) had them. Assuming a Poisson world with constant demand rates, the inventory investment of $193,717,400 would be expected to produce an operational availability of 12.9%, increasing to 55.8% with cannibalization. The study team noted that the 55.8% rate was very close to the actual C-5 fleet availability at that time, and that high cannibalization rates were a major problem then.

When the initial stock was optimally redistributed, the operational availability increased to 20.2%, or 56.2% with cannibalization. Third, an optimal run was made with a 70% aircraft availability target (92.6% with cannibalization). This required a budget of only $56,782,960. Finally, another optimal run was made with a 95% aircraft availability target with a resulting budget of $138.197m.

In summary, the study predicted an operational availability (without cannibalization) of 95% at a cost of $138.197m using the optimal VARI-METRIC theory as compared to an availability of 12.9% at a cost of $193.717m under Readiness Based Sparing levels.[1]

10.7 Model Implementation - Army

The Army uses VARI-METRIC theory in a model called SESAME - Selected Essential-Item Stockage for Availability MEthod, originally documented in Kaplan (1980). SESAME is used for initial spares budgeting and for item procurement in an effectively four echelon, multi-indenture calculation. Since 1993, it has been required by the Army for all initial provisioning. The model incorporates base differences so that the Army is able to distribute the initial stocks to the appropriate bases. Private contractors doing business with the Army are supposed to use a multi-echelon approach also.

SESAME is scheduled to be used for follow-on provisioning by 2006. The delay until now has been organizational in nature, where some material was owned by the units and some by the logistics organization. Now all material is coming under the stock fund for management. There is also a dynamic form of SESAME which is used for wartime planning.

The Army has not had the Air Force problems with DLA, but they are trying to figure out ways to coordinate the DLA computed stocks with the Army requirements. They are also trying to use coordinated or distributed decision-making to control the amount of turbulence in stock levels that can come about because of new programs.[2]

[1] Private communication with Jim Russell, TFD Group, Monterey, CA.
[2] Private communication with Meyer Kotkin, Army Material Command (AMC).

10.8 Model Implementation - Navy

The Navy has used a model called ACIM, Availability Centered Inventory Model, developed by Clark (1981). Although the mathematics is similar to that of MOD-METRIC, Clark did some of the earliest multi-echelon inventory research, Clark and Scarf (1960), and he was one of the very first researchers to relate inventory policy to availability in an optimal model.

The Navy runs ACIM for the wholesale level first, obtaining depot reponse times to the users. More often a standard depot time is used. Then a multi-indenture, single echelon version of the model is run to determine retail stock levels. Of course, the optimal split of budget between wholesale and retail is unknown, so the model run in two parts can not be optimal.

Another model called SEASCAPE is used on the AEGIS weapon system. This model, built by RCA, incorporates some redundancy modeling, but is used at the retail level only.

The aviation community uses a model called ARROWS - Aviation Readiness Related Operation Weapon Systems. This model is used for initial provisioning (allowance lists) in a multi-indenture, single echelon version. There is also a Spares Budgeting Model. All of the models use pieces of the theory described in this book.

10.9 Model Implementation – Coast Guard

The Coast Guard has been using VARI-METRIC theory as embodied in the VMetric model (see Appendix E) for a number of years. In a study of the use of VMetric at an actual site, Air Station Cape Cod, their Research and Development report (Rodriguez et al. 1997) concluded that overall savings to the Coast Guard could exceed $200M. In general, what they have learned and accepted is how single item sparing significantly misrepresents the required sparing where a tie to system availability can be made, as it can with VMetric, and generally requires too many spares of the wrong parts. So, while one could show big savings, much of that is consumed by adjusting the domain of the spares to include the parts not spared in sufficient quantity, since they can't recapitalize excesses. They have used VMetric for all four types of their aircraft to redistribute those they already have for maximum availability, and have reset allowances accordingly. At present, they use the model to set their strategic stocking levels for all aircraft at all sites and their depot, and regularly conduct analyses of scenarios arising out of the routine management of a system that large. VMetric is an integral part of their over arching logistics management system, ALMIS, and it will be used to provide the initial sparing for the introduction of the C130J into the Coast Guard fleet. It is most likely that VMetric will remain their sparing model for the foreseeable future.[1]

10.10 Model Implementation - Worldwide

Some years ago we visited a nuclear power company and briefed a version of VARI-METRIC. When we explained that they could reduce their investment in spares, the logistics manager said, "Why would we care about that? We just tell the regulatory commission what we spent and then our rates are set to reimburse us." Fortunately, that attitude is seldom heard anymore.

The basic VARI-METRIC theory is embodied in other computer programs such as OPUS, developed in Sweden. This multi-echelon, multi-indenture model is used extensively in Europe, by aircraft manufacturers and several NATO air forces. VARI-METRIC theory has been used by hundreds of manufacturers in the United States, particularly those involved with aircraft and related systems. In most cases the application is initial stockage lists, but it is hard to know the full range of model usage. As noted in Chapters 6 and 7, VARI-METRIC theory was used extensively by NASA in Space Station Freedom.

We should mention an important commercial application of multi-echelon theory for managing spare parts at IBM described in Cohen et al. (1990), in which (s, *s*) theory is applied in a four-echelon system with over 200,000 part numbers. The measure of effectiveness is fill rate with different targets at different echelons. The authors note that the system improved performance in several ways: (1) improvements in demand forecasting; (2) accounting for the multi-echelon structure; (3) accounting for part commonality; (4) enhancing cost-service tradeoffs. The focus of the article is to minimize costs subject to service constraints, where the costs include replenishment cost, expedited delivery costs, and inventory holding cost, (s, *s*) theory is required because many of the items have high demand. It is estimated that average inventory investment to achieve the same level of service was reduced by 20 to 25 percent - a saving of $0.25 billion.

10.11 Model Hierarchies

We have talked about the engineering parts hierarchy many times in this book, and have noted that stock levels should vary by indenture. There is another important kind of hierarchy that we need to address here, and this is

[1] Private communication from Capt. Norman Scurria, USCG ret., now an employee of TFD Group, Monterey, CA.

a kind of hierarchy of models based on when the models are used during the life of a system.

At the earliest conceptual stages of designing a new system, crude models may be used to make various design trade-offs. Even the word "model" may be glorifying some of the engineering trade-offs that are made early on. When models are used at this stage they are likely to be deterministic. For example, in the space station it is important to have a rough idea of the remove-and-replace workload long before the detailed system design. Obviously if the workload is inconsistent with astronaut hours available, major redesign is required.

Later as the designs are finalized, we can begin to use optimal sparing models such as VARI-METRIC. Even when there is a lot of uncertainty about item demand rates and costs, it makes sense to use the model for budgeting purposes. Over time the input data should become more reliable, and procurement decisions can be made; but, the Air Force realizes a procurement model that is run today will not deliver spares until after a procurement lead time. Rather than attempt to predict the exact configuration of aircraft at each base and the flying hour program sometime in the future, they run the procurement model on the assumption that all bases are identical. This simplifies both the data estimation and computation problems.

When spares are delivered, the base differences must be taken into account. As we argued in Chapter 8 when we discussed the DRIVE model, the one thing we know at the end of the procurement lead time is that we didn't buy the right stuff. Regardless of the method chosen for buying spares, it turns out that because of changing demand rates, condemnations, and experience, the world is somewhat different than we had planned a lead time earlier.

That is, of course, why a model such as DRIVE exists - to take the latest information and use our repair resources and distribution capabilities to achieve the best short-run performance possible with the assets in the system. That is why we want to consider the age of installed items that are subject to wear out when we make the individual shuttle manifests for the space station.

Although this book has been primarily concerned with the long-range optimal spares procurement problems, it is important to consider the hierarchy of different models that are required for proper management: from deterministic, aggregate models in the early stages to very detailed, operational models for day-to-day management at the later stages.

10.12 System Approach Revisited One More Time

This book has employed the system approach extensively. We saw in Chapter 1 that this is not just a different perspective than the traditional item approach, but that it can lead to significantly different policies and substantial improvements in cost and performance. We have demonstrated this with numerous examples using actual data wherever possible.

We noted that the models developed in this book are all analytic. Simulation is used in a couple of cases to verify the accuracy of the analytic models, but the models themselves are mathematical equations which can be solved for optimal stockage policies in an efficient manner. The analytic nature of the models is essential for practical implementation on personal computers or even main frames.

Because we have used the system approach in so many different ways throughout the book, we want to remind the reader of how it has been employed:

1. Generation of cost-availability curves for a system

2. Determination of the optimal stock level for an item at a location, taking into account that it depends on stock levels at other locations and for other items, as well as considering the target availability or cost

3. Estimation of mean demand for an item (Bayes or James-Stein techniques that utilize data from other items)

4. Estimation of item variance-to-mean ratios from a power curve or Bayesian techniques (using data from other items)

5. Evaluation of prediction techniques (using a stockage model and availability for a group of items)

The item approach was simpler, because stockage decisions on each item were made independently. But we never knew whether those item decisions would produce an acceptable availability level or whether the budget requirements were feasible. Furthermore, the item approach used the same protection level on every item regardless of unit cost, indenture, echelon, and cannibalization policy. We have seen that dramatic improvements in system performance and cost are possible when these are taken into account systematically.

10.13 Problems

1. Suppose that there is a piece of test equipment that could be provided to each base that would enable the base to repair more first-indenture items. The cost of the test equipment has to be weighed against the benefit. Consider the cost as fixed and determine whether the benefit of the test equipment at bases increases or decreases with each of the following: (1) higher target availabilities; (2) longer repair times; (3) longer order-and-ship times; (4) longer program life; (5) reduction in unit cost; (6) increase in demand rates; (7) no increase in demand rates but an increase in variability; (8) more failures in the test equipment itself; (9) higher skill requirements for operators; (10) more false positives and negatives in the equipment diagnoses (11) easier cannibalizations.

2. In Section 9.7 we noted that the Navy sometimes uses the model theory twice: (1) to establish resupply times from depot in a wholesale model; (2) to establish retail stock levels based on depot resupply times. Does this procedure lead to an optimal solution?

3. The system approach can be applied to other logistics problems. Consider the problem of determining retention levels for stock in long supply as modeled by Kaplan (1969). The problem is to compare the cost of keeping an item for n years where a benefit is received at that time versus disposing of the asset now at a fraction of its cost.

We define:

c = cost of the item
f = fraction of item cost from disposal sale
F = fraction of item cost for storage annually
i = discount rate for future costs/benefits

We assume that there is demand history enabling us to compute approximately the annual demand rate. If we dispose of a unit of the item now, we receive a benefit, fc. If we use the unit n years from now, we obtain a benefit (discounted for the fact that it is n years in the future) less a cost due to the storage every year from now to year n which is

$$c/(1+i)^n - cF\sum_{j=1}^{n} 1/(1+i)^j$$

The objective is to compute the largest value of n (the retention limit) for which the benefit of keeping the unit for n years before it is used exceeds the benefit from immediate disposal.

Note that the purchase cost c is a sunk cost. But we assume that the value of the item if used now is c, or the cost if reprocured now is c. Since f and F are expressed as fractions, c itself drops out of the calculation. Assume that $i = .10$ and $F = .01$, and show for disposal fractions f of 0, .05, .1, .15. and .2 that the optimal number of years for retention, n, beyond which units should be sold is 25, 20.5, 17.5, 15.5, and 13.5 years respectively.

The system approach comes about in the determination of the stockage cost fraction F. When warehouses are less than 85% full, the annual cost of stocking a slow-moving item is very small. Beyond 85%, the warehouse incurs additional charges for moving inventory to make room for new receipts, and there is a greater chance of being unable to locate the item when it is demanded. As the warehouse becomes full, the cost of new storage relates to the imputed cost of an additional warehouse. Thus, the cost of storage per unit volume as a function of warehouse capacity might be depicted as in Figure 10-1.

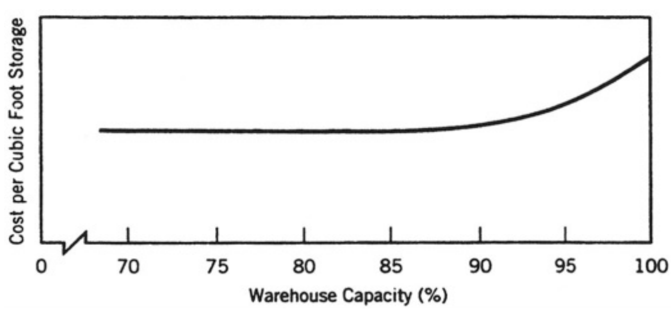


Figure 10-1. **Cost of storage per cubic foot as a function of warehouse capacity**

4. Suppose that the items in Problem 3 can deteriorate during storage, decreasing their value. What modification needs to be made in the equation?

Appendix A

PALM'S THEOREM

If a man will begin with certainties he shall end with doubts, but if he will be content to begin with doubts he shall end in certainties.

-Sir Francis Bacon

A.1 Appendix Overview

In this appendix we prove the standard version of Palm's theorem as stated in Chapter 2 using the approach in Hadley and Whitin (1963). The theorem has been extended in several other references. Feeney and Sherbrooke (1966) showed that if demand is compound Poisson with mean m and the repair time for each failed unit is independently and identically distributed according to any distribution with mean repair time T, then the steady-state probability distribution for the number of units in repair is the same compound Poisson distribution with rate mT, provided that the time drawn from the repair time distribution is the same for all demands in a compounding "cluster". It is shown there that Palm's theorem can be extended to cover the lost sales case also, where the probabilities are defined for zero to some maximum number of items in repair, s. The probabilities are the same compound Poisson probabilities, normalized by the sum of the probabilities from 0 to s.

When Feeney and Sherbrooke (1966) was published, we thought the compound Poisson model might be a better representation than the Poisson of the repairable item demand process. As indicated elsewhere in this book, we believe now that the variance-to-mean ratios greater than one that we observe are because of Poisson processes with changing means. The results in that article do apply to Poisson demand and the lost sales. But since we are not interested in compound Poisson demand or the lost sales case in this book, the article will not be discussed below.

Palm's theorem has been extended to finite calling populations by Sherbrooke (1966), and the steady-state probabilities are shown below in Section A.4.

The most important extension of Palm's theorem was by Crawford (1981) who proved a dynamic version that applies when the Poisson demand rate (and even the repair distribution) is changing over time. Although it was not explicitly noted, we have been using this result in Chapter 4 and elsewhere when we talk about Poisson processes with changing means.

A.2 Preliminary Mathematics

Before proving Palm's theorem, we need to demonstrate a property of the Poisson distribution. The joint probability that n Poisson events occur in a time interval t such that the first occurs between t_1 and $t_1 + dt_1$, the second between t_2 and $t_2 + dt_2$, etc. and the nth between t_n and $t_n + dt_n$ where $t_1 < t_2 \ldots < t_n$ is:

$$[e^{-mt_1} m dt_1][e^{-m(t_2-t_1)} m dt_2]\ldots[e^{-m(t_n-t_{n-1})} m dt_n][e^{-m(t-t_n)}]$$

which simplifies to

$$m^n e^{-mt} dt_1 dt_2 \ldots dt_n \tag{A.1}$$

The conditional probability of n events occurring as outlined above is the joint probability in Equation A.1 divided by the Poisson probability of n events which yields

$$[n!/t^n] dt_1 dt_2 \ldots dt_n \tag{A.2}$$

The proof that Equation A.2 is a legitimate density function is left to Problem 1. Note that the density function is independent of the t_i, so that the occurrences of the different events are independent random variables. The probability that an event occurs between t_i and $t_i + dt_i$ is just dt_i/t.

Let's consider how the $n!$ comes about. Imagine the time interval from 0 to t divided into $2n + 1$ boxes, n of which correspond to the intervals t_i to $t_i +$

dt_i, and the remaining $n + 1$ correspond to the remaining $n + 1$ intervals. Think of the n Poisson events as balls that must be tossed into the boxes corresponding to the intervals t_i to $t_i + dt_i$. The probability that one of the balls goes into the box t_1 to $t_1 + dt_1$ is ndt_1/t. The probability that one of the remaining balls goes into the box t_2 to $t_2 + dt_2$ is $(n - 1)dt_2/t$, etc.

We have shown that if we know that at least one Poisson event has occurred in the time interval from 0 to t, then the probability that any one occurs between τ and $\tau + d\tau$ is $d\tau/t$, and this is independent of how many events have occurred in the interval and their times of occurrence.

A.3 Proof of Palm's Theorem

PALM'S THEOREM. If demand for an item is given by a Poisson process with mean m per unit time and the repair time for each failed unit is independently and identically distributed according to any distribution with mean repair time T, then the steady-state probability distribution for the number of units in repair has a Poisson distribution with mean mT.

PROOF: Let $h(\tau)$ be the probability that the repair time is τ, where the mean repair time is T. The probability that a unit entering repair at time t_i will finish by time $t > t_i$ is $H(t - t_i)$

$$H(t-t_i) = \int_0^{t-t_i} h(\tau)d\tau \tag{A.3}$$

If we know that at least one demand has occurred in the time interval 0 to t, the probability that the unit is repaired by time t is

$$[H(t-t_i)/t]dt_i$$

Thus the probability that any particular demand occurring in the interval 0 to t is repaired by t is

$$[1/t]\int_0^t H(t-t_i)dt_i = [1/t]\int_0^t H(\tau)d\tau \tag{A.4}$$

Suppose that were u units on hand initially and let x be the *net inventory*, defined as the on-hand minus backorders. The basic idea of the proof is to let the time t go to infinity and show that the number of units in repair has a Poisson distribution with mean mT.

We want to compute the probability that the net inventory is x at time t if there have been y demands since the system began operation. Now x is a number that ranges from a maximum of u to a minimum of $u - y$ (in the case where no repairs have been completed by time t), so that $u - x$ represents the number of units still in repair. Since there have been y demands, there must have been $y + x - u$ completed repairs. The probability that $u - x$ of the y demands have been repaired is given by a binomial distribution:

$$\text{bin}(x) = \binom{y}{u-x}[(1-t)\int_0^t H(\tau)d\tau]^{y+x-u}[(1/t)\int_0^t \{1-H(\tau)\}d\tau]^{u-x}$$

$$0 \leq x \leq y \qquad \text{(A.5)}$$

The probability in Equation A.5 is conditional on the number of demands, y. Weighting by this Poisson probability and summing over all $y \geq u - x$ yields the unconditional probability that the net inventory is x at time t

$$\Pr\{\text{net inventory is } x \text{ at time } t\} = \sum_{y=u-x}^{\infty} p(y \mid mt)bin(x)$$

$$= [1/(u-x)!][m\int_0^t \{1-H(\tau)\}d\tau]e^{-m\int_0^t\{1-H(\tau)\}d\tau} \qquad \text{(A.6)}$$

To estimate the limit of Pr{net inventory is x at time t} as $t \rightarrow \infty$, we note

$$\lim_{t\to\infty} \int_0^t \{1-H(\tau)\}d\tau = \int_0^\infty \{1-H(\tau)\}d\tau = \int_0^\infty -\tau d[1-H(\tau)]d\tau$$

$$= \int_0^\infty \tau(dH/d\tau)d\tau = \int_0^\infty \tau h(\tau)d\tau = T \qquad \text{(A.7)}$$

where the last step on the upper line is obtained using integration by parts. When Equation A.7 is substituted into Equation A.6 we obtain

$$\lim_{t\to\infty} \Pr\{\text{net inventory is } x \text{ at time } t\} = p(u - x|mT)$$

$$x = u, u - 1, u - 2 \ldots \qquad \text{(A.8)}$$

Since u - x is the number of units in repair with a range from 0 to ∞, Palm's theorem is proved.

A.4 Extension of Palm's Theorem to Finite Populations

PALM'S THEOREM FOR A FINITE CALLING POPULATION. Consider a population of s units where demand for each unit is given by a Poisson process with mean m per unit time and the repair time for each unit is independently and identically distributed according to any distribution with mean repair time T. Then the steady-state probability distribution for the number in repair is given by

$$h(y) = \binom{N}{y}(mT)^y / D \tag{A.9}$$

where D is a normalizing constant.

It is easy to derive this result for exponential or constant repair times using the standard birth and death equations as in Feller (1958). The proof for arbitrary repair time distributions in Sherbrooke (1966) is a variation on the very complicated derivation in Feeney and Sherbrooke (1966). Since this result is never used in the book, it seems inappropriate to do more than state it here.

A.5 Dynamic Form of Palm's Theorem

DYNAMIC FORM OF PALM'S THEOREM: Suppose that the demand process is Poisson with mean function $\lambda(\tau)$ defined for $\tau > 0$. Assume that a demand occurring at time τ has a probability of not being repaired by time $t > \tau$ given by $\overline{H}(\tau,t)$, which is independent of the times at which any other demands occur. Then the number of units in repair at time t is a Poisson random variable with mean $m(t)$ given by

$$m(t) = \int_0^t \overline{H}(\tau,t)\lambda(\tau)d\tau \tag{A.10}$$

We have simplified the statement of the theorem in Crawford (1981) omitting some of the mathematical niceties of measure theory. In effect, we assume that the integral defined above exists. We have used the notation $\overline{H}$ as it is the complement of the function H that we used in Section A.3.

While it is possible for $\overline{H}(\tau,t)$ to be a function of τ and t, it is usually assumed in most applications that the repair time distribution is a function of the difference $(t - \tau)$ only, and thus that the survival time distribution $\overline{H}(\tau,t) = \overline{H}(t-\tau)$ where $\tau < t$. In other words the repair time distribution is stationary. A particularly simple case is when the repair time is constant, say T. In that case $m(t)$ is just the integral of $\lambda(\tau)$ from t - T to T.

Note that if the demand rate is constant so that $\lambda(\tau) = m$, $\overline{H}(\tau,t) = \overline{H}(t-\tau)$, and we let t go to infinity, then Equation A.10 simplifies to the usual form of Palm's theorem:

$$\lim_{t \to \infty} m \int_0^t \{1 - H(t-\tau)\} d\tau = mT$$

where the last equality follows from Equations A.4 and A.7.

In view of the independence between the demand process and the repair process, the dynamic version of Palm's theorem is highly plausible. Since the proof in Crawford is rather lengthy and involves some advanced mathematical arguments, we will not prove it here.

A.6 Problems

1. Show that Equation A.2 is a proper probability density function by integrating over each t_i where $0 < t_1 < t_2 ... < t_n < t$.

2. Do you think that the Dynamic Form of Palm's Theorem can be extended to cover finite populations? If so, restate the theorem in Section A.5. Remember that the statement of the theorem must make sense in the special case of stationary demand.

3. Expediting – In Sherbrooke (1971) it is shown that Palm's theorem still applies when demand rates or repair rates are a function of the number of units in repair. The steady-state probabilities are not Poisson, but they depend only on the mean of the repair time distribution - not its shape. The key assumption is that any modification to the repair rate (such as a speed-up to account for expediting) must affect all units in repair. Equation A.9 is a special case of this result. It is of limited use in logistics modeling, because the decrease in demand due to an aircraft down can be caused by any first-indenture backorder. This makes the problem non-separable.

Consider a different type of expediting and an item that is never repaired at the operating site. There is a standard order-and-ship time from the depot, but when the base has a backorder, an expedited order-and-ship time is used instead. Assume Poisson demand and use the birth and death equation approach in Feller (1958) to derive the steady-state probabilities for exponential service. Explain why this result can or cannot be extended to arbitrary repair time distributions.

Write a computer program for your solution. Suppose that the pipeline is 1, the stock level is 2, and that with expediting the pipeline can be cut in half. Show that the expected backorders of .1036 from Table 4.2 can be cut to .0663 if expediting is invoked whenever there is a backorder or to .0370 if expediting is invoked whenever the stock on-hand drops to zero.

Appendix B

MULTI-ECHELON SYSTEMS WITH LATERAL SUPPLY

I do not fear computers, I fear the lack of them -Isaac Asimov

B.1 Appendix Overview

We develop approximations to estimate the expected backorders in a multi-echelon system in which lateral supply actions between bases are allowed when a first-indenture backorder occurs. These approximations are easy to compute, and the average absolute error over a wide range of parameter values is less than 4% when items are depot-repairable, even when bases are dissimilar. With lateral supply, backorder reductions of 30% to 50% are not uncommon, and a 72% reduction was observed in two cases. Lateral supply becomes more important with low demand rates.

A similar approach was unsuccessful for base-repairable items. Lateral supply has a beneficial effect only when the lateral supply time is very short, one-fourth or less of the average base repair time. Even in such cases lateral supply is unlikely to be important in an actual application, because base management can expedite repair of critical items.

There are two major reasons for including this appendix: The first is qualitative - the results provide some insight as to the conditions under which lateral supply should provide significant reductions in backorders. The second is methodological - even though we have been unable to build an analytic model for lateral supply, we have illustrated a statistical approach using simulation results. This enables us to represent the impact of lateral supply in our models such as VMetric (see Appendix E).

B.2 Background

The military services have become more interested recently in the topic of lateral supply between bases to reduce first-indenture backorders, "holes" in end items such as aircraft. In the case of the U.S. Air Force, this is due to information system improvements and physical capabilities such as the European Distribution System and a similar system in the Pacific. These systems have regularly scheduled flights that can make a lateral supply in a couple of days. In December 1988 alone, the Air Force in Europe responded to over 1500 redistribution requests for material needed by other locations worldwide.

A lateral supply is made whenever a demand at a base causes a backorder (i.e., stock on hand is zero and a customer has an unfilled demand) and a spare on hand at some other base can be lateraled to arrive at the base with the backorder before an item already in transit from a depot or completing base repair. In other words, a lateral supply is not made if the laterally supplied spare would arrive after the backorder has been satisfied. Backorders are of interest, because the logarithm of availability is proportional to the sum of base backorders as shown in Section 2.14.

On the modeling side, there have been several recent papers. Slay (1986) described an approach to solving this problem for identical bases at a Multi-Echelon Inventory Conference. He includes the possibility of "delayed lateral" shipments: A delayed lateral supply may be made when a backorder cannot be satisfied by an immediate lateral supply because no base has stock, but some base receives stock at a later point in time and can laterally supply the deficient base before a spare can be provided by another source. Our simulations of the problem show that these delayed lateral actions are important; they can exceed the immediate lateral actions when system stock is quite low, and are more important for systems of many bases.

Slay derives complex formulas for the lower and upper bounds for expected backorders (EBOs). We programmed his solution, and compared it with a simulation. On a sample of five cases, the average absolute error of his technique was about 21%; his upper bounds were closer with an average absolute error of 9%, though inappropriately labelled since they were exceeded by the simulated solution in all but one case. This motivated us to consider simpler approaches to estimating lower and upper bounds that would be more accurate and applicable to cases where bases are not identical.

Lee (1987) presents another approach as well as a comprehensive review of earlier research. He considers clusters of similar bases, with laterals made within each cluster. Axsater (1990) attacks the same problem with different techniques resulting in smaller errors than Lee when compared to simulation results. The Axsater techniques have the advantage that they can be applied to dissimilar bases as well. However, both authors assume that laterals are made from a randomly chosen base in the group that has stock on hand, and neither makes provision for delayed laterals (though Axsater recognizes their potential importance). Since these decision rules are not optimal, we did not program them.

Our objective is to find an accurate approximation of the EBOs under lateral supply that is easy to compute, capable of handling dissimilar bases, and capable of handling cases with some base repair In addition we have attempted to understand the conditions under which lateral supply is likely to make a significant reduction in backorders. It turns out that when the amount of stock is so low that there are always multiple backorders, lateral supply is of little help; similarly, when stock is very high, lateral actions are rarely needed.

Culosi in Groover, et al. (1987) noticed that bounds on the backorders under lateral supply are easily obtained using standard multi-echelon models such as VARI-METRIC, described in Chapter 5, to optimally allocate between the depot and bases any amount of spare stock. A lower bound on backorders can be estimated by combining all bases into a single base, because this is equivalent to instantaneous lateral supply between bases; an upper bound using the actual configuration of bases and demand rates, because VARI-METRIC allocates stock between bases and depot to minimize base backorders when there is no lateral supply. Culosi used some of the Slay analytic relations to develop interpolation formulas between the bounds.

We have adopted the Culosi idea of using interpolation between lower and upper bounds from VARI-METRIC, but with a statistical approach based on regression. We have run a large number of simulations to check the accuracy of the formulas, and the results for depot-repairable items have been very encouraging. Furthermore, the computation is fast and accurate even for different types of bases. The Culosi approach is not described in detail here because it is based on the extensive derivations in Slay, and because the results are uniformly less accurate than our regression results.

B.3 Simulation Description

A few comments about our simulation are in order since it is used as the standard for comparing the lateral approximations. We consider a single item and a system consisting of several bases at which demands occur and a supporting depot. We use the METRIC assumptions of Chapter 3, the most significant of which is that no queuing occurs for repair (the infinite channel queue assumption). When the repair times for each unit are independent and demand is Poisson, the steady-state probability distribution for the number of units in repair is Poisson by Palm's theorem, depending only on the mean of the repair distribution.

The simulation can be run with either constant or exponential repair times, and similar results are obtained as expected from Palm's theorem. Confidence intervals are computed for the simulated EBOs, based on the EBO autocorrelation functions described in standard texts such as Gross and Harris (1974). Since the confidence intervals around the simulated backorders tend to be shorter for constant repair times, they were used in most cases. We assume that the status of all units of stock (in transit between specific sites, in repair at a specific site, or on the shelf at a specific site) is known at any point. Furthermore, we assume that we know the time at which each shipment will arrive. Demand at each base is assumed to be Poisson distributed with a known, constant mean (although the rates may vary from one base to another). The percentage of demand that is base-repairable is known, and the balance is assumed to be depot-repairable. Base and depot average repair times are known.

The numerical results presented below are satisfactory only for the case of depot-repairable items, so our further discussion of the simulation will be confined to them. Because repairable items tend to have high cost and low demand at a base, the economic order quantity is one. Thus a resupply request is placed on the depot on a one-for-one basis whenever there is a demand, and the unserviceable unit is shipped to the depot for repair. In a system with no lateral supply, the stock level at each base equals the stock on hand at the base plus the number in resupply to the base minus backorders. The stock level remains constant, but its composition varies probabilistically over time. When there is a demand, the stock on hand, if positive, decreases by one or the number of backorders increases by one, and this is balanced by an increase of one in the number of resupply requests; when a resupply is received at a base, the number of backorders, if positive, is decreased by one or the stock on hand is increased by one, and this is balanced by a decrease of one in the number of outstanding resupply requests.

Thus, the number of resupply requests from a base at any point in time is related to the base stock level. The VARI-METRIC model assumes that resupply requests are filled on a first-come, first-served basis. However, in a lateral supply system with perfect information about the location of serviceable and unserviceable units, decisions on where to ship from and to are made on the basis of expected need as defined below. Thus, the concept of a stock level has little relevance.

Simulation results show that when lateral shipment times are less than the shipment time from a depot to a base, any item completing depot repair should be sent to some base immediately. When all bases are identical, the appropriate base to receive a unit of stock from depot is the one with the smallest value for the *inventory position,* defined as (stock on hand) + (stock already in-transit to it from the depot or being lateraled from another base) minus backorders. Similarly, when some base incurs a backorder, the appropriate base to send a lateral is that base with positive stock on hand that has the largest value for the inventory position.

When bases are different, we must subtract expected base demand over a planning horizon from the above quantities. For shipments from depot to the bases, we use a depot planning horizon that is about twice the length of the depot-to-base order and ship time. For laterals between bases we use a base planning horizon that is about 1.5 times as long as the lateral supply time. The results are not extremely sensitive to either planning horizon length. We would expect that the planning horizon should be at least the order-and-ship time in the first case and the lateral supply time in the second, because those are the earliest times that the system can be affected by those decisions. The fact that the best planning horizons are somewhat longer is an empirical observation for which we have no theoretical basis.

The lateral supply times that we use in the simulation, motivated by the Air Force application, are typically only a couple of days. Under the assumptions of accurate information, it is unlikely that we would want to divert a lateral shipment that is already underway. In other applications of lateral supply with longer times, this capability may be important.

B.4 Parameter Values

We examined a number of cases in which parameters reflected U.S. Air Force experience. Our numbers and the approximate mean and standard deviation for the Air Force are shown in Table B-1. The order and ship time is defined to be the time interval from the placement of a base resupply request on the depot until the resupply arrives at the base, given that the depot had stock on hand.

Our data source did not give the mean and standard deviation for the number of bases or the percentage of items that are base-repairable, but our parameter choices cover the range of interest. The distributions for the remaining four parameters tend to be skewed with long right-hand tails. Consequently, we have not chosen parameter values symmetrically around the means.

Table B-1. Range of Parameter Values for U.S. Air Force

Measure	Mean	Std.Dev.
Number of bases (2, 4, 5, 10, 15, 20, and 30)	NA	NA
Order-and-ship time average (8, 16, and 30 days)	16	4
Depot repair time average (30, 60, and 120 days)	60	30
Daily demand rate average at base (0.011, 0.022, 0.055, 0.11, 0.22, 0.44, 0.88 units/day)	.04	.30
Base repair percent average (0, 50, and 100)	NA	NA
Base repair time average (4, 8, and 16 days)	7	4

Source: Air Force Recoverable Item Requirements System (D041). Statistics from several D041 data bases, from 1984–1986. NA, not applicable.

B.5 Depot-Repairable-Only Items

The case of depot repair is particularly important because the time to obtain resupply from the depot can be fairly long, particularly if there are no spares on the shelf when the request is received at depot. In contrast, if the item is base-repairable and a backorder exists (an aircraft is down for the item), base management can often expedite the repair and preclude the need for lateral supply.

In Table B-2, 16 different parameter value combinations are listed: the number of bases, the average daily demand rate at each base, the average order-and-ship time (days) from depot, the average depot repair time (days), and the average pipeline (defined as the average number of units of stock in repair or resupply at a random point in time). The pipeline is a measurable quantity, and can be computed as the sum over all bases of

Average
daily demand x (order and ship time + depot repair time).
per base

In Case 1 of Table B-2 this equals 5 x 0.88 x 38 = 167.2. The retrograde shipment time from base to depot of a broken item is by convention included in the depot repair time. The parameter a in the last column is the estimated value of the dependent variable in a regression that is used in Equation B.3, as described below.

The cases in Table B-2 are presented in order of decreasing demand. For a given demand level, the cases are in order of decreasing number of bases. The depot repair times always exceed the order-and-ship times, as is typical, though a range of values is included. At the end of the table, several cases with dissimilar bases are presented. Cases 11 and 13-15 are comprised of two base types each and Case 16 has three base types.

The cases in Table B-3 are the parameter values from Table B-2 combined with a lateral supply time and the number of repairable units of stock in the system. The parameter values and the number of units of stock are used to compute a lower bound, LB, on the total backorders at bases when all bases are combined into one base (equivalent to a lateral supply ime of zero); to compute an upper bound, UB, when stock is optimally allocated to the bases and depot by VARI-METRIC under the assumption of no lateral supply. The *estimated* backorders, EB, are obtained from the lower and upper bounds and an estimating relationship described below, which utilizes the lateral supply time as well.

Table B-2. **Depot-Repairable Parameters**

Case	Number of Bases	Daily Demand Rate	Order- & Ship-Days	Depot-Repair Days	Average Pipeline	Regression Estimate a
1	5	0.88	8	30	167.2	.426
2	20	0.22	8	30	167.2	.268
3	10	0.22	8	30	83.6	.268
4	5	0.22	30	120	165.0	.129
5	10	0.22	16	30	101.2	.183
6	10	0.22	8	60	149.6	.268
7	2	0.22	8	30	16.7	.268
8	5	0.22	8	30	41.8	.268
9	5	0.22	16	120	149.6	.183
10	5	0.055	8	30	10.5	.169
11	5	0.22	8	30	62.7	.243
	5	0.11	8	30		
12	5	0.055	16	60	20.9	.115
13	2	0.44	8	30	41.8	.289
	2	0.11	8	30		
14	5	0.44	16	120	448.8	.183
	10	0.11	16	120		
15	20	0.11	8	60	187.0	.200
	10	0.055	8	60		
16	5	0.055	8	30	16.7	.136
	5	0.022	8	30		
	5	0.011	8	30		

Table B-3. Expected Backorders under Lateral Supply (Depot Repairable)

			Model Backorders			Simulated			
Case	Stk	T	LB	UB	EB	EBO	LPipe	% Err.	% Red.
1a	160	2	9.504	11.659	10.739	10.824	0.902	−0.8	7.2
6a	140	2	11.069	13.461	12.060	12.020	0.908	0.3	10.7
8a	30	2	11.863	12.218	12.010	12.066	0.117	−0.5	1.2
b	40	2	3.556	4.893	4.110	4.160	0.469	−1.2	15.0
c	40	8	3.556	4.893	4.736	4.620	0.369	0.1	2.7
11a	60	2	4.665	7.351	5.699	5.800	1.023	−1.7	21.1
13a	36	2	6.408	7.089	6.707	6.690	0.275	0.3	5.6
	LESS THAN PIPELINE : AVERAGE ABSOLUTE							0.7	9.2
1b	180	2	1.137	2.931	2.164	2.217	0.521	−2.4	24.4
d	200	4	0.031	0.377	0.313	0.306	0.017	2.4	18.8
3a	100	1	0.156	2.013	0.592	0.566	0.358	4.6	71.9
c	100	4	0.156	2.013	1.376	1.421	0.459	−3.2	29.4
e	120	2	0.001	0.234	0.097	0.100	0.059	−2.6	57.3
4b	182	8	0.592	1.935	1.455	1.448	0.360	0.5	25.2
5b	104	4	2.782	7.798	5.378	5.745	2.125	−6.4*	26.3
d	120	2	0.142	2.781	0.948	0.926	0.583	2.3	66.7
f	140	2	0.001	0.595	0.182	0.168	0.109	8.4*	71.8
6b	170	4	0.269	1.820	1.288	1.240	0.460	3.9	31.9
d	190	4	0.002	0.258	0.171	0.159	0.041	7.3	38.5
8d	45	2	1.313	2.599	1.846	1.838	0.440	0.4	29.3
f	50	2	0.349	1.267	0.730	0.724	0.280	0.8	42.8
9b	180	2	0.033	0.216	0.089	0.090	0.038	−1.2	58.3
11b	80	2	0.054	1.213	0.500	0.507	0.344	−1.3	58.2
12a	25	2	0.496	1.341	0.669	0.651	0.157	2.8	51.5
13c	52	2	0.187	0.654	0.392	0.398	0.137	−1.6	39.2
14a	480	4	0.696	4.504	2.667	2.740	1.469	−2.7	39.2
15a	205	2	0.638	6.780	2.661	2.764	1.809	−3.7	59.2
16a	20	2	0.521	2.241	0.932	0.958	0.418	−2.7	57.3
	REGRESSION SET : AVERAGE ABSOLUTE							3.1	44.8
1c	200	2	0.031	0.377	0.229	0.237	0.058	−3.4	37.1
2a	200	2	0.031	3.427	1.438	1.404	0.875	2.5	59.0
3b	100	2	0.156	2.013	0.926	0.934	0.548	−0.9	53.6
d	111	2	0.006	0.688	0.289	0.271	0.168	6.7*	60.6
4a	180	4	0.808	2.342	1.425	1.423	0.365	0.1	39.2
5a	104	2	2.782	7.798	4.314	4.560	1.441	−5.4*	41.5
c	110	4	1.095	5.341	3.292	3.460	1.570	−4.8*	35.2
e	120	4	0.142	2.781	1.507	1.570	0.780	−4.0	43.5

Table B-3. **(Continued)**

			Model Backorders			Simulated			
Case	Stk	T	LB	UB	EB	EBO	LPipe	% Err.	% Red.
g	140	4	0.001	0.595	0.308	0.295	0.121	4.5*	50.4
6c	190	2	0.002	0.258	0.108	0.104	0.053	4.3	59.7
7a	20	2	0.521	0.758	0.619	0.624	0.072	−0.7	17.7
8e	45	8	1.313	2.599	2.448	2.360	0.215	3.7	9.2
9a	160	4	1.381	2.648	2.037	2.050	0.396	−0.6	22.6
10a	15	2	0.147	0.668	0.296	0.291	0.144	1.7	56.5
11c	90	2	0.001	0.375	0.145	0.155	0.098	-6.3*	58.7
13b	44	2	1.644	2.656	2.087	2.080	0.322	0.3	21.7
d	52	4	0.187	0.654	0.507	0.509	0.161	−0.4	22.2
14b	487	4	0.328	3.332	1.883	2.026	1.113	−7.1	39.2
15b	217	2	0.083	3.799	1.307	1.435	1.093	−8.9*	62.2
16b	22	2	0.222	1.947	0.635	0.649	0.372	−2.2	66.7
	EVALUATION SET : AVERAGE ABSOLUTE							3.4	42.8

Stk, Stock; T, lateral supply time in days; LB, lower bound; UB, upper bound; EB, estimated backorders; EBO, expected backorders from simulation; LPipe, lateral supply pipeline between bases; % Err, % error; % Red., % reduction in backorders due to lateral supply; *, estimate is outside 95% confidence interval.

For example, the average pipeline in Case 16, using Equation B.1,

$$= 5 \text{ x } (0.055 + 0.022 + 0.011) \text{ x } 38 = 16.72$$

The lower bound on expected backorders, LB, for Case 16a with a stock level s = 20 in Table B-3 is

$$\sum_{x>s}(x-s)p(x) = .0521 \tag{B.2}$$

where $p(x)$ is the Poisson probability of x units in resupply with a mean of 16.72.

The upper bound is more complicated. The depot pipeline is the depot demand times the depot repair time. Since all base demands are repaired at depot this equals

$$5 \text{ x } (0.055 + 0.022 + 0.011) \text{ x } 30 = 13.2.$$

The pipeline to a large base with mean 0.055 is the mean demand of 0.055 multiplied by (the order-and-ship time plus the average delay due to the fact that the depot does not always have stock on the shelf); when the depot stock level is 0 this equals 0.055 x 38 = 2.09, but the pipeline decreases as the depot stock level increases. Starting with system stock of zero, the VARI-METRIC model determines the optimal allocation of each unit of stock to some base or the depot as described in Chapter 5.

The simulation results in Table B-3 are shown for the expected backorders, EBO, and the expected number of units in the lateral supply pipeline between bases, LPipe. The backorders from the regression model are compared with the simulated backorders using the percent error, % Err. as 100 x (EB - EBO)/(EBO). The percent reduction in backorders due to lateral supply, % Red., is 100 x (UB - EBO)/EBO.

The first seven cases in Table B-3 have less stock than the pipeline, which accounts for the small reduction in backorders due to lateral supply. The remaining 40 cases were split into two groups of 20, and regression was used on the first group to estimate the parameters of an estimating relationship. Then the adequacy of the relationship was evaluated or *calibrated* by using it to predict the backorders of the second group of 20.

The objective of the estimating relationship is to determine f, an interpolation function with values between 0 and 1 that can be applied to the difference between the upper bound, UB, and the lower bound, LB, to obtain an estimate for backorders, EB:

$$\mathrm{EB} = \mathrm{LB} + f(\mathrm{UB} - \mathrm{LB}) \qquad \text{(B.3)}$$

Reviewing the simulation results, we noted that the value of f had an exponential relationship with T when all other parameter values were held constant. Thus, instead of including T as a variable in the regression with the need to estimate another coefficient, we decided to represent f by the following equation:

$$f = 1 - e^{-aT}) \qquad 0 \le \mathrm{T} < O \qquad \text{(B.4)}$$

where:

f = interpolation value

T = lateral supply time in days (assumed to be a fixed constant that is the same between any two bases)

a = parameter estimated by regression, described below

O = order and shipping time from depot to base (time in days to receive an order from depot when the depot has stock on hand)

With this formulation, $T = 0$ results in $f = 0$, which is appropriate because the several bases have become a single base with EBOs equal to the lower bound. An infinite value for T corresponds to no lateral supply for which a value of f = 1, corresponding to the upper bound, is appropriate. We will restrict the upper value for T to the order-and-ship time, O, because we are not really interested in lateral supply times that exceed O.

The regression relation for a was[1]:

$$a = 1.406 O^{-0.554} D^{0.334} \tag{B.5}$$

where:

D = sum of the daily demand rates across bases/number of bases

It was noted above that the value of a is independent of the amount of stock and the lateral supply time; happily we have found here that the value of a is independent of the depot repair time and the number of bases, also. The estimated value of a from the regression is shown in the last column of Table B-2.

For example, when the order and ship time of 8 days and the average base demand rate of 0.22 demands/day of Case 3 are substituted into Equation B.5, the value of a is found to be 0.268 as indicated in Table B-2. When Equation B.3 is plotted for a =.268 and combined with the lower and upper bounds for Cases 3a-3c of Table B-3, corresponding to total stock of 100, Figure B-1 is obtained. The three simulated backorder solutions for T = 1, 2, and 4 are shown in Table B-4 as well.

***Table B-4.* Three Simulated Backorder Solutions for T = 1, 2, and 4**

Lateral Supply Time (T) in Days	Estimated Backorders (EB)	Expected Backorders (EBO)
1	.592	.566
2	.926	.934
4	1.376	1.421

Equation for estimated backorders is $.156 + 1.857\,(1 - e^{-.268T})$

[1] When Equations B.3 and B.4 are solved for A the result is:

a = -(1/T)log[(UB-EB)/(UB-LB)].

The "true value" of a for each case in Table B-3 is imputed from this equation by substituting the simulated actual backorders for EB Then multiple linear regression was applied to the logarithmic form of Equation B.5. The only significant independent variables were O and D, and each of the three coefficients had a t-value with magnitude greater than 11, indicating a probability less than .00005 that the coefficient is not significant. Regression has been used to estimate functions of interest to logisticians by many authors, e.g. Ehrhardt(1979).

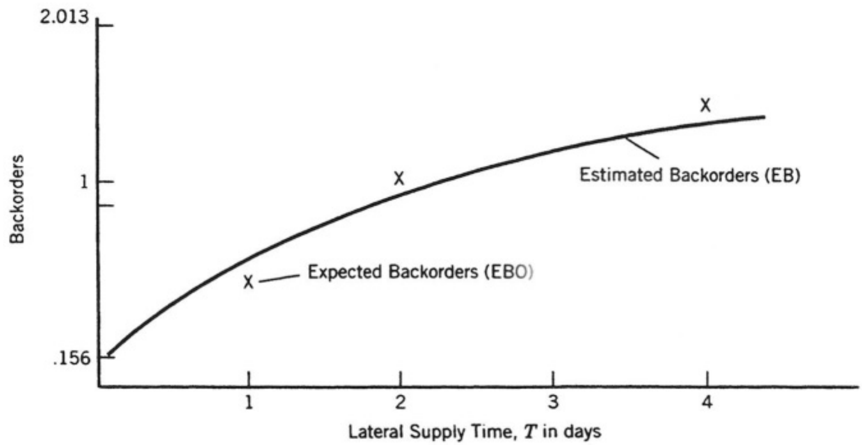


Figure B-1. **Comparison of estimated and actual backorders for Cases 3a-3c.**

As noted above, the first 7 cases in Table B-3 are of limited interest because the stock in the system is less than the pipeline. Their average absolute error of 0.7% is very low, but the reduction in backorders due to lateral supply is very low as well at 9.2%.

The next 20 cases in Table B-3 were used to estimate the regression coefficients. Although a wide range of parameter values was used, the average absolute error for the 20 cases is only 3.1%. The two asterisks after the percentage error for Cases 5b and 5f are used to indicate that the estimated EBO are outside the 95% confidence interval of the simulated EBO. The largest percentage errors in this group are 8.4% for Case 5f and 7.3% for Case 6d, but the absolute errors are quite small in both cases. For these 20 cases the average percentage reduction in EBOs due to lateral supply is 44.8%, but this value would increase with more cases having shorter lateral supply times and vice-versa.

The last 20 cases in Table B-3 were used to evaluate the accuracy of the regression coefficients. Their average absolute error increases to only 3.4%, though there are now six cases marked by asterisks where the estimate lies outside the 95% confidence limits from the simulation. The largest percentage errors are -8.9% for Case 15b and 6.7% for Case 3d, comparable to the largest errors in the regression set. The average percent reduction in backorders of 42.8% is only slightly less for this group of 20 cases.

The parameters for the 20 cases in the regression set were chosen to represent the range of cases of interest. This is why the set of 20 cases in the evaluation set has similar parameter values. However, we would expect little degradation with different parameter values because there are only three regression coefficients.

It is interesting to note that very large reductions in backorders are possible with lateral supply, even though the average number of units in the lateral supply pipeline tends to be modest. The largest percentage reductions of 71.9% and 71.8% in Table B-3 had an average of only of .358 and .109 units respectively in the lateral pipeline. Since the total stock was 100 and 140 in the two cases, it is clear that a little lateral supply can go a long way. The largest number of units in lateral supply was 2.125, only 2% of the stock in that case.

In summary, the results for depot-repairable-only items are very good. A wide range of parameter values was simulated, and the degradation of the regression from the prediction set to the evaluation set was modest. It is clear that lateral supply can result in dramatic reductions in EBOs. The simplicity of the calculation suggests that it should be useful.

B.6 Base-Repairable Items

Results for items that are fully (or partially) base-repairable were unsatisfactory. When the lateral supply time is held constant for a group of cases sharing the same parameters, the value of a changes as stock is added. This is because all stock is at the bases, and if bases are identical, the VARI-METRIC upper bound tends to be lower when each base has the same stock level. When the total stock does not divide evenly into the number of bases, the backorders with no lateral supply are larger. Thus, lateral supply gives an even bigger improvement over VARI-METRIC, which has no lateral supply, when the spares do not divide equally.

In the depot-repairable cases with identical bases, this did not happen. As stock is added to the system, VARI-METRIC puts the same stock level at each base and adjusts the amount of depot stock.

Because of the variation in a, we used regression for the interpolation value f itself. Although the coefficients were all statistically significant, five independent variables were needed instead of two, and the average error was much larger. In more than half of the cases, the errors were outside the 95% confidence limits. The maximum error was 14.3%, and one absolute error was quite large as well (0.609). For these reasons we have not bothered to even present the regression equations.

We ran 13 cases where the lateral supply time was one-half or more of the base repair time. In 5 cases the simulated backorders exceeded the upper bound; in the other 8 cases the simulated backorders were only slightly less. We conclude that lateral supply will have a beneficial impact only when it is one-fourth or less of the base repair time.

Even if the lateral supply time is one-fourth or less of the base repair time, the benefits of lateral supply are probably overstated. As noted above, the base can change its repair priorities and repair an item, if it has repair parts, in less than the average repair time. Thus, lateral supply will provide little benefit when items are base-repairable.

B.7 Number of Lateral Shipments

It is desirable to estimate the number of units in the lateral pipeline (both immediate and delayed), shown in the column headed LPipe in Table B-3. We used step-wise regression for the depot-repairable-only cases. The only statistically significant independent variables were: (1) estimated backorders under lateral supply; (2) the percentage reduction in backorders due to lateral supply. However, the average absolute error of nearly 20% on both the regression set and the evaluation set was so large that the results are not worth presenting.

B.8 Summary

Exact analytic solutions to many logistics problems can be obtained only by making unrealistic simplifying assumptions. An alternative is to construct efficient computational techniques that provide close approximations to the solutions of the real problems over a large range of simulated situations. We have chosen to take that approach in this analysis of a long-standing and difficult problem

We have demonstrated simple computational models that provide estimates for the EBOs when lateral supply between bases is allowed. The estimates have an absolute average error of less than 4% for depot-repairable-only items, even when bases are dissimilar.

The benefit of lateral supply is negligible unless the spares in the system exceed the pipeline, and lateral supply is rarely needed when spares are more than 1.5 times the pipeline. However, for intermediate spares values, lateral supply provides large improvements in both absolute and percentage terms. Backorder reductions of 30% to 50% are not uncommon, and a 72% reduction was observed in two cases. Lateral supply has a greater impact when the demand rates are low.

When items are base-repairable or partially base-repairable, a similar approximation procedure for estimating EBOs under lateral supply was unsuccessful. Furthermore, unless the lateral supply time is one-fourth or less of the base repair time, lateral supply actions may degrade performance. Even if the lateral supply time is one-fourth or less of the base repair time, the benefits of lateral supply in an application are probably overstated. This is because a base can change its repair priorities and repair an item, if it has repair parts, in less than the average repair time, rather than resort to a lateral.

Appendix C

DEMAND PREDICTION STUDIES

If the facts don't fit the theory, change the facts. -Albert Einstein

C.1 Background

Demand prediction for an item consists of two essential elements: (1) predicting the mean demand; (2) predicting the variance-to-mean ratio. The problem is difficult, because the mean demand may vary with time, and the variance-to-mean ratio increases with the length of the time period over which it is measured.

For a number of years the Air Force has used an 8-quarter moving average to estimate the mean demand over the next quarter. If an estimate of demand is required for an arbitrary period of T years, the quarterly estimate is multiplied by $4T$. Recently, Sherbrooke (1984) found that for predicting quarterly demand for two years in the future, exponential smoothing with a constant of 0.4 applied to quarterly data could produce a 39% reduction in squared error and a 12% reduction in average absolute error. The data were 1,027 items selected as a stratified random sample from the Air Force Logistics Command (AFLC) inventory.

The variance-to-mean ratio is important, because it is used in the determination of how much safety stock to buy on an item. The Air Force uses the assumption of Poisson demand in many stockage models, implying a variance-to-mean ratio of one. This simplifies the demand prediction problem, but makes the unrealistic assumption that the demand rate is constant over time. As we saw in Table 4.4, the Poisson assumption tends to misallocate investment resulting in low availabilities.

Stevens and Hill (1973) developed an estimating relationship for the variance-to-mean ratio, *V*, which has been altered by the Air Force Logistics Command:

$$\hat{V} = 1.132(\hat{m}T)^{.3407} \tag{C.1}$$

where $\hat{m}T$ is the estimate of the average pipeline (the average number of units in repair or resupply). The ^ indicates an estimate. In the original research the independent variable was the average annual demand, *m,* rather than *mT* and we will see below that the distinction is significant.

Sherbrooke (1984) noted that in the Stevens-Hill estimation of this formula, 27% of the items were excluded because their demand was not "sufficiently stable". As a result the formula understates the true variability of the item population. The authors estimated mean demand as the average of eight quarters, but did not consider alternatives. This is important because good estimates of the mean are critical to making good estimates of variance-to mean ratio. Also, the authors did not consider the use of program element data, such as flying hours. Finally, Stevens and Hill estimated a variance-to-mean ratio relationship, but they never evaluated its performance as we shall do below.

Using a slightly different functional form so that the variance-to-mean ratio can never be less than one, and including all 1,027 items with any positive demand history Sherbrooke (1984) found the best relationship to be

$$\hat{V} = 1 + .14\hat{m}^{.58} \tag{C.2}$$

The results in this appendix are drawn from a more recent study by Sherbrooke (1987) using data from several aircraft types. The advantage of having data on all first-indenture items for an aircraft type is that it is possible to use availability as the measure of performance. This is more meaningful than such measures as squared error, percent error, or average absolute error where the errors on different items are weighted equally, because availability is the Air Force's objective when it allocates funds to spare parts.

Another objective of the study was to see whether the best technique for predicting the mean and variance-to-mean ratio of demand is similar from one aircraft type to another. Moreover, though quarterly Recoverable Consumption Item Requirements (D041) data serve as the primary source as in the earlier study, it was important to obtain some transaction data from an individual base. This information gives the day on which each demand occurred, and thus tells us more about the "physics" of the demand process.

In Chapter 4 we described briefly some of the major results of these demand prediction studies. We found that exponential smoothing with a smoothing constant of 0.4 on quarterly data was the best predictor of the mean, and that a slightly different power curve relationship than Equation C.2 was the best estimator for the variance-to-mean ratio, *V,* over the next year:

$$\hat{V} = 1 + .14\hat{m}^{.5} \tag{C.3}$$

where $\hat{m}$ is the estimate of the annual demand and $\hat{V}$ is constrained to be no more than 20. The experimental procedure was summarized in Figure 4.3.

We realize that many readers are concerned with applications that may differ substantially from the Air Force. In other problems exponential smoothing may not be the best predictor of the mean, and the best estimating relationship for variance-to-mean ratios could be quite different from Equation C.3. But, we believe that the approach used in this appendix for determining the best prediction technique has wide applicability to other logistics problems as well.

Since the completion of the study described in this appendix, another empirical analysis by Slay and Sherbrooke (1988) was performed that provided more evidence that demand over short periods of time does follow a Poisson process with a constant mean. The larger variance-to-mean ratios observed over longer periods of time arise because the mean of the demand process changes leading to a Poisson process with non-stationary increments.

C.2 Appendix Overview

In this chapter we want to provide more detail on the demand prediction studies that have been performed than was possible in Chapter 4. Section C.3 below describes the demand prediction experiment methodology, including the justification for using an availability model to evaluate the accuracy of each technique (a "technique" being defined as a process for predicting the mean demand and the variance-to-mean ratio).

The candidate demand prediction techniques are defined in Section C.4, and results with the C-5 airframe presented. Sections C.5 and C.6 present similar analyses for the A-10 airframe and the F-16 airframe/engine. For each system, the same prediction technique appears to provide the best results.

For the F-16 the flying-hour program increased dramatically over the 16 quarters of history. We would expect that demand/flying hour would be more stable than demand/quarter for many items, and this was found to be true. On an item basis the estimator with less variability over the 12 quarter base period is selected, and its performance against an Air Force procedure based on Equation C.1 is described in Section C.7.

In Section C.8 we examine correlational information from all of the aircraft programs above and some A-10 daily demand data from England AFB. All of these data show that demand in a quarter is mostly highly correlated with demand in the neighboring quarter, and that the correlation decreases as the distance between quarters lengthens. This provides further support for any prediction technique, such as exponential smoothing, that weights more recent data more heavily.

When a smoothing constant of 0.4 is applied to quarterly data, only 13% of the weight is applied to demands that occurred more than a year before (as contrasted with 50% for an 8 quarter moving average). It has been suggested that for very low-demand items, it may be better to use a smaller smoothing constant to obtain more history. Section C.9 uses the data on each aircraft type to refute this hypothesis.

Section C.10 provides the conclusions from all of our work on demand prediction. This includes methodological recommendations on how demand prediction studies should be performed and practical recommendations for logisticians.

We have created a Windows tool, the Demand Analysis System (DAS) described in Appendix F, that does all the analyses described in this chapter.

C.3 Description of the Demand Prediction Experiment

The objectives of this study are to find demand prediction techniques that: (1) attain a high level of system availability in the future and (2) predict a level of system availability similar to the level attained.

The candidate procedures for predicting mean demand fall into the four categories shown in Table C-1. Let $\hat{d}$ be the predicted mean quarterly demand, d_1 the demand in the most recent quarter, d_2 the demand in the preceding quarter, etc. For exponential smoothing, an example is shown where the smoothing constant, ω, is set at 0.4. The Air Force now uses the second procedure, a moving average, with N = 8. This weights each of the last 8 quarters equally and gives no weight to earlier quarters.

Table C-1. **Procedures for Predicting Mean Demand**

1. **Exponential smoothing**

$$\hat{d} = \omega d_1 + \omega(1 - \omega)d_2 + \omega(1 - \omega)^2 d_3 + \cdots$$

Example: $\omega = .4$

$$\hat{d} = .4d_1 + .24d_2 + .144d_3 \cdots$$

2. **Moving average**

$$\hat{d} = (d_1 + d_2 + \cdots + d_n)/n$$

Example: $n = 8$

$$\hat{d} = .125d_1 + .125d_2 + \cdots + .125d_8$$

3. **Bayes**

Prediction is a combination of observed demand and the prior probability distribution. Estimate of a variance is provided.

4. **Higher-order procedures with trend also estimated**

The Bayesian procedure is the "objective Bayes" technique that was applied at George Air Force Base in 1965 and described in Section 4.10. The basic idea is that the two parameters of a single gamma prior probability distribution are estimated from past demand on all items. Bayes theorem is then used to combine this prior distribution with specific data about an item to obtain posterior distributions that will differ by item. An important difference between the Bayes procedure and the two previous procedures is that the Bayesian approach yields a probability distribution - instead of a point estimate - of demand. There is, therefore, no need to develop a variance-estimating procedure for the Bayes technique.

Finally, it is possible to use higher-order procedures where both a trend and a mean value are estimated. We followed a procedure called "Holt linear estimation," which is described in Makridakis and Hibon (1979); it consists basically of second-order exponential smoothing. The first parameter is the same as the first-order exponential smoothing parameter and is set at 0.4, as above. The second parameter is a trend value that we set at 0.5 after some trial and error. However, we restricted the rate of growth for trend to eliminate absurd inferences or estimates of negative demand; the absolute value of the trend/quarter was limited to 15% of the mean value, so that over 4 periods the mean could increase or decrease by 60% at most.

Earlier work (Sherbrooke 1984) made it clear that it is important to estimate the variance-to-mean ratio, *V,* as well as the mean. Four types of candidate procedures are considered in Table C-2. The first is a power relation, where the variance-to-mean ratio is an increasing function of the mean. Here the mean is expressed as an annual value. If the mean is expressed over any other time period, the exponent is unchanged, but the multiplicative constant would be altered.

Another procedure is to estimate the item's variance from its own past data. A third type of procedure is the Bayesian, which automatically provides an estimate of variance. Finally, recent research, as in Winkler and Makridakis (1983), suggests that combinations of procedures may outperform individual procedures. A general power relation such as Procedure 1 in combination with Procedure 2, which is sensitive to extreme variation in the history of an item, may outperform either procedure alone. To prevent any technique from making absurdly large estimates of variance-to-mean ratio, we invoked an arbitrary maximum of $V= 20$ on each item.

Table C-2. **Procedures for Predicting Variance-to-Mean Ratio**

1. **Power relations.** ***Example:*** $\hat{V} = 1 + .14\hat{m}^{.5}$
2. **Based on historical data for the item**
3. **Bayesian**
4. **Combinations of procedures 1, 2, and 3**

A major difference between this study and earlier studies of demand is the evaluation mechanism. Most prediction studies use some model-independent measure to evaluate accuracy. The problem is that there are many candidates, and it is hard to decide which, if any, is appropriate. Average absolute error is a common measure, but should the error on a very high demand item be divided by some factor before adding it to the error on a low-demand item? Should the same error on two items that are widely different in unit costs be weighted equally? If the error is twice as large, should the penalty be twice as heavy - or should it be heavier as in a squared error criterion? Is it reasonable to assign as much weight to an over-prediction error as to an under-prediction error? Many economic studies use percentage error, but many of our items have a "true demand" of zero. This causes problems as a divisor.

Our view is that aircraft availability should be used, because the objective of demand prediction on a group of items is to maximize availability for a specified budget. The formula for availability was derived in Equation 2.18, and is reproduced in Table C-3 below where we have assumed that the quantity of each item in the aircraft is one, and the number of aircraft is 100.

Table C-3. **Evaluation of Demand Prediction Techniques**

Model-independent measures	
	Average absolute error
	Mean-squared error
	Percentage error
Model-dependent measure	
	Availability $= 100 \prod_{i}^{I} [1 - \mathrm{EBO}_i(s_i)/100]$

The data for this demand prediction experiment were taken from 16 quarters of D041 data. We drew on the first 12 quarters for historical data and used the last 4 quarters to evaluate the accuracy of each prediction technique. The data used in the calculations are listed in Table C-4. We assume that each demand is repairable at the base or at the depot; there are no condemnations. The average repair time by item is the sum of: (1) the average fraction base repairable multiplied by the base repair time and (2) the average fraction depot repairable multiplied by: (the order-and-ship time plus the average depot delay caused by not having an item on the shelf). We assumed an order and ship time of 2 weeks. The average depot delay is unknown, but it will tend to be longer if the depot repair time is longer. As an arbitrary but reasonable procedure, we, assumed that the depot delay would be 0.3 times the depot repair time. (We also used 0.1 with qualitatively similar results.)

The prediction process for each technique, summarized in Table C-4, consists of taking the 12 historical quarters of demand data and estimating the mean and variance-to-mean ratio of demand for each item. The mean demand over the repair time and the variance-to-mean ratio are used to estimate the two parameters of a negative binomial distribution of demand. This demand distribution is used to calculate the backorders as a function of the stock level for each item. Then these backorder functions are used with unit cost to marginally allocate the fixed budget optimally across all first-indenture items on the aircraft. (This is precisely what we did with marginal analysis in Chapter 2). By using the same budget for each prediction technique, we can compare performance of the techniques.

The result is a set of stock levels and a predicted availability for each technique. In the case of the Holt second-order smoothing which predicts trend, stock levels from the marginal allocation will be different for each of the four predicted quarters.

We evaluated the prediction techniques as summarized in Table C-5. The inputs for the evaluation are the actual demands during the last 4 quarters, not used for prediction, and the stock levels by item. Unfortunately, we do not know the day on which each demand occurred, only the total quarterly demand. The only option is to use a random number generator to draw the day during the quarter for each demand. Of course, the random assignment of demands across a quarter may smooth out the real, unknown daily demands, but most of the demand fluctuation is captured with the four quarterly totals. On the basis of the item data on repair times, we compute the day on which repair of the item will be completed. (The average repair time for the item is used, not a draw from a probability distribution with that mean.) The same sequence of demands and repair times is used for all techniques.

Table C-4. Demand Prediction Process

A. Item Input
   1. Demand data for each of 12 quarters
   2. Average fraction of demands that is base-repairable
   3. Average base repair time
   4. Average depot repair time
   5. Unit cost

B. System Input
   1. Budget to allocate across items
   2. Specification of several demand prediction techniques (procedure for estimating the mean and variance-to-mean of every item)

C. Process
   1. Calculate expected back orders as a function of stock level for each prediction technique and item
   2. Use marginal analysis to allocate investment across items to maximize the predicted availability using a specific prediction technique and the budget

D. Output by Prediction Technique
   1. Stock level for each item
   2. Predicted availability

Table C-5. Evaluation Process

A. Input
   1. Total demand by item for each of 4 prediction quarters
   2. Item stock levels by technique (and quarter when the technique estimates the trend)

B. Process
   1. Draw from a uniform distribution the day (during the quarter) on which each demand occurs
   2. Using the item's average repair time, determine the day that each repair is completed

C. Output
   Average attained availability by technique

The output of this activity is an "attained availability" for each technique. We have put quotes around this availability because it was attained in a simulation of the last year, not actually observed in the field. The availability is computed daily and averaged for each of the 4 quarters and for the entire year. Availability with cannibalization, consolidation of shortages onto the smallest number of aircraft, is also computed.

C.4 Results of the Demand Prediction Experiment for C-5 Airframe

Each of the next three sections provides the results of the demand prediction experiments for a particular aircraft system: C-5 airframe, A-10 airframe, and F-16 engine/airframe respectively. A separate section is devoted to each aircraft type, because the demand prediction techniques differed somewhat. In the case of the F-16 there were dramatic increases in the flying program over the sixteen quarters. Consequently, it was necessary to investigate whether demand per flying hour was a better predictor than demand per quarter.

One problem with D041 quarterly demand data is that all demand at bases is aggregated. For our demand prediction experiments it would be preferable to have demand data by base. As our first aircraft system, we selected the C-5, a large transport aircraft operated by the Military Airlift Command, because it operates from only two home bases. Worldwide demand, as drawn from the D041, is therefore not too different from base level demand, although demands do occur at other en route bases during flight. All non-engine, first-indenture items (line-replaceable-units) with some demand were used. There were 560 such items on the C-5 airframe.

We selected 17 candidate techniques in Table C-6 that had shown promise in earlier studies. By a "technique", we mean a combination of a procedure for estimating the annual mean and a procedure for estimating the variance-to-mean ratio. For example, the first four techniques in Table C-6 use the same "objective" Bayesian procedure described in Section 4.10 to estimate the variance-to-mean ratio, but different procedures to estimate the mean itself.

One departure from the Bayes procedure in Section 4.10 is used for items with more than 1,200 demands annually during the first 12 quarters of demand history. For the 39 C-5 items with this property, the Bayesian procedure underestimates significantly the true mean demand (because it combines the information on all 560 items). For these items, we apply Technique 11 which produces the best results in Table C-7 ($80 million budget) and Table C-8 ($100 million budget).

Technique 5 is the constant, Poisson assumption that the variance-to-mean ratio equals one; Technique 6 is a constant variance-to-mean ratio of 3. Techniques 7 through 11 use power curve relationships for the variance-to-mean ratio based on the estimate of annual demand, *m,* whereas techniques 12 through 14 are based on estimates of the average pipeline, *mT*.

Table C-6. List of Demand Prediction Techniques

Technique	Mean Demand	$\hat{V}$
1	First-order exponential smoothing using .3	Bayes
2	First-order exponential smoothing using .4	Bayes
3	First-order exponential smoothing using 1. (uses last quarter of history only)	Bayes
4	8-quarter moving average	Bayes
5	First-order exponential smoothing using .4	1
6	First-order exponential smoothing using .4	3
7	First-order exponential smoothing using .4	$1 + .14\hat{m}^{.58}$
8	8-quarter moving average	$1 + .14\hat{m}^{.58}$
9	First-order exponential smoothing using .4	$1 + .216\hat{m}^{.69}$
10	First-order exponential smoothing using .4	$1 + .10\hat{m}^{.58}$
11	First-order exponential smoothing using .4	$1 + .14\hat{m}^{.5}$
12	First-order exponential smoothing using .4	$1 + .8\hat{m}T$
13	First-roder exponential smoothing using .4	$1 + 1.8(\hat{m}T)^{.58}$
14	First-order exponential smoothing using .4	$2 + 1.8(\hat{m}T)^{.58}$
15	First-order exponential smoothing using .4	From item data
16	First-order exponential smoothing using .4	Tech. 7 and 15
17	Holt linear exponential smoothing .4 and .5	From item data

V, variance-to-mean ratio; m, mean annual demand; mT, average pipeline.

The parameter values for Techniques 7-14 are variations on values obtained by nonlinear estimation applied to Air Force Logistics Command (AFLC) data in Sherbrooke (1984). We will see below that Techniques 7 - 11 tend to perform better than Techniques 12 - 14. A detailed comparison of the two approaches is presented below in Section C.7.

Technique 15 estimates the variance-to-mean ratio for the item from the 12 quarters of demand history on the item. Technique 16 applies the results of recent forecasting literature where the suggestion is made that forecasts derived from combinations of techniques often outperform the individual techniques.

Technique 17 is the only one that attempts to predict trend in the mean demand. It combines exponential smoothing with a second trend parameter with the limit described above on the rate of increase or decrease per quarter to 15% of the mean demand. The equations are not presented here because the results, as described below, were extremely poor.

As noted in the previous section, the value of each demand prediction technique is assessed by two availabilities: the attained availability and its relationship to the predicted availability. In each table of results we show the attained availability first, because a high value during the prediction year is the more important consideration. A less important but desirable property is to have a predicted availability that is fairly close to the attained availability.

***Table C-7.* Availability of C-5 Airframe: $80 million budget**

		Availability (%)					
	Cost	Attained	Predicted	Attained by Quarter			
Technique	($M)	Annual	Annual	1	2	3	4
1	80.08	81.2	85.6	82.7	85.6	85.1	71.4
2	80.07	82.5	84.6	85.3	87.4	85.8	71.5
3	80.00	70.1	83.8	85.3	67.8	69.8	57.6
4	80.06	68.6	89.1	68.7	70.4	72.3	63.0
5	80.00	56.6	98.6	84.9	76.5	63.1	2.0
6	80.18	67.8	75.8	87.0	85.0	83.3	16.0
7	80.23	81.0	64.1	85.6	84.1	85.1	69.0
8	80.00	79.6	60.9	82.6	81.8	87.1	66.9
9	80.13	74.6	41.9	74.2	78.1	81.6	64.5
10	80.05	82.2	74.0	88.6	84.2	84.8	71.3
11*	80.14	82.8*	77.1	89.1	85.2	85.1	71.9
12	80.01	75.2	73.1	74.3	79.0	82.8	64.7
13	80.00	75.6	72.4	75.2	79.1	82.8	65.2
14	80.01	71.0	67.7	69.8	74.4	78.0	61.8
15	80.03	77.6	63.9	78.9	81.0	84.8	65.8
16	80.02	79.5	60.6	81.2	83.4	86.9	66.7
17	80.01	69.3	70.7	73.7	73.2	75.8	54.3

*** Highest attained availability for budget of $80 million.**

Our choice is Technique 11, marked with asterisks in Tables C.7 and C.8, because it has the highest attained availability for a budget of $80 million and the fourth highest availability for a budget of $100 million. The three Bayesian techniques that finish ahead of Technique 11 in Table C-8 are overoptimistic; their predicted availabilities always exceed the attained. Technique 11 produces predicted availabilities that are reasonably close to the attained values. The Bayesian techniques have the additional disadvantages of being significantly more complex and unsuitable in an initial support planning context, because they require some observed demand data to estimate variance-to-mean ratios; the other techniques require only an estimate of the mean. Also, the Bayes techniques would have done substantially worse if we had not altered the predictions for items with more than 1,200 demands per year during the 12 quarters of history. Note that the results for Techniques 1 and 2 are quite similar, indicating a degree of robustness to the smoothing constant value. However, the smoothing constant of 1 for technique 3 produces distinctly inferior results. This is not too surprising, because this is equivalent to a Markov process assumption where demand in the last quarter contains all useful information.

Table C-8. Availability of C-5 Airframe: $100 Million Budget

		Availability (%)					
	Cost	Attained	Predicted	Attained by Quarter			
Technique	($M)	Annual	Annual	1	2	3	4
1	100.19	92.5	94.6	98.1	97.4	95.7	78.7
2	100.05	91.7	94.3	97.1	96.5	94.6	78.7
3	100.01	79.0	95.2	92.6	77.5	77.1	68.7
4	100.07	92.4	95.8	94.5	97.6	96.2	81.3
5	100.07	62.2	99.9	91.9	85.8	68.9	2.3
6	100.02	77.4	89.3	96.4	94.0	92.4	26.7
7	100.00	90.5	82.0	95.2	93.0	92.8	81.2
8	100.04	90.5	79.8	95.1	93.2	95.8	78.1
9	100.00	88.9	61.2	92.9	92.0	94.0	76.5
10	100.02	91.3	89.2	96.4	93.7	92.5	82.8
11*	100.18	91.3*	91.1	96.8	93.4	91.9	83.1
12	100.00	88.8	84.4	92.3	92.1	94.4	76.5
13	100.03	89.6	83.8	93.4	93.0	94.9	77.1
14	100.02	87.8	79.6	91.7	90.9	93.5	75.0
15	100.07	89.3	81.2	93.0	93.5	94.7	75.9
16	100.00	90.6	79.0	94.8	94.2	95.8	77.4
17	100.04	83.1	88.0	86.6	86.5	89.2	70.3

*** Fourth highest availability for budget of $100 million.**

A few comments should be made about the other techniques. Technique 5, the Poisson assumption, yields especially poor attained availabilities results, particularly as we move to predicted quarter 4; Technique 6, with a constant variance-to-mean ratio of 3, produces better though similar results. Not surprisingly, Techniques 7-11, as a group produce similar results. It is surprising to note that the 8 quarter moving average, Techniques 4 and 8, is quite inferior to exponential smoothing techniques with the same variance-to-mean ratio procedure at a budget of $80 million; quite similar at a budget of $100 million. It is a little unexpected that Technique 9 did not outperform Technique 11, because nonlinear estimation techniques were used to estimate the best relationship for the C-5 from its own data.

Techniques 12-14 use a variance-to-mean ratio based on the estimate of the average pipeline. The poor performance of these techniques was a little hard to understand at first. Since the mean repair/resupply times vary by item from 6 to 57 days, the values of the average pipeline, mT, are much smaller than the item annual demand, m. Accordingly, larger coefficients are needed for comparable variability. However, it appears that these techniques are inferior to Technique 11, because it is not the variability over the repair/resupply time that is important, but the drift in the mean demand rate over the year.

The mean annual demand influences the variance-to-mean ratio of Techniques 7-14, but Techniques 12-14 depend on the mean repair time, T, as well. Table C-9 compares the variance-to-mean ratio as a function of repair time for the best technique in each group, as well as that for Equation C.1 ($\hat{V} = 1.132(\hat{m}T)^{.3407}$) which is considered in Section C.7 below. As noted earlier the maximum variance-to-mean ratio is set at 20, and a minimum of 1 for Equation C.1.

Technique 15 reinforces the results of other studies that suggest the item variance-to-mean ratios estimated from past history on the item tend to be very unstable, particularly since we believe the item means are drifting over time. Technique 16 does not seem to be an improvement over its components, Techniques 7 and 15. Finally, the attempts to predict trend in the mean value from Technique 17 are consistent with results reported in Sherbrooke (1984). In both Tables C.7 and C.8 the trend technique gives results that worsen with time, the opposite of what we might hope. Our explanation is that the trend technique, and second-order smoothing, suffer because of "regression to the mean"; it is highly unlikely that a series of data with a significant trend will continue indefinitely. It is like predicting a larger rainfall next year when we have just had a 50-year flood.

Though demand means tend to drift over time, we have been unable to predict the direction of this drift for individual items.

Table C-9. **Variance-to-Mean Ratio Over Repair Time**

m	Technique 11 (Any Repair time)	Technique 13 $T = 6$	Technique 13 $T = 30$	Technique 13 $T = 57$	Equation C.1 $T = 6$	Equation C.1 $T = 30$	Equation C.1 $T = 57$
10	1.44	1.63	2.61	3.33	1.00	1.06	1.32
100	2.40	3.40	7.11	9.86	1.34	2.32	2.89
1000	5.40	10.13	20.00	20.00	2.94	5.08	6.33

C.5 Results of the Demand Prediction Experiment for A-10 Airframe

Similar comparisons were made on the 480 items of the A-10 airframe. Ten of the more promising techniques from the 17 in Table C-6 were chosen; the results re given in Table C-10. Technique 11M is like Technique 11 of Table C-6, except that the mean is estimated with the 8 quarter moving average. Since both techniques employ the same variance-to-mean ratio relationship, this enables us to compare the moving average and exponential smoothing procedures for predicting the mean.

Table C-10. **Availability of A-10 Airframe: $80 Million Budget**

		Availability (%)					
				Attained by Quarter			Attained
Technique	Cpst ($M)	Attained Annual	Predicted Annual	1	2	3	Attained Annual (with Cannibal.)
1	80.00	89.2	100.0	72.6	97.6	97.3	89.4
2	80.01	92.7	100.0	82.0	97.5	98.3	92.9
3	80.05	57.3	100.0	40.6	95.2	36.1	57.8
4	80.00	87.0	100.0	80.1	85.2	95.7	90.6
5	80.07	96.6	97.9	90.4	99.9	99.4	96.6
11M	80.00	70.3	98.8	50.0	72.9	88.0	79.0
11*	80.15	97.7*	97.2	99.8	96.3	97.0	98.0
13	80.02	93.1	61.7	95.8	94.5	89.0	96.1
15	80.01	86.9	73.5	94.1	84.5	82.1	93.0
17	80.08	76.4	60.4	91.4	72.5	65.3	91.5

* **Best availability.**

Technique 11 is the best technique for the A-10 aircraft, indicated by the asterisks in Table C-10, confirming the results for the C-5 aircraft found earlier. It gives the highest attained availability, 97.7% (only 3 quarters are shown); it yields the average availability of 98.0% in the last column when full cannibalization, consolidation of shortages on the fewest aircraft possible, is practiced. Technique 11M (moving average) is far inferior to Technique 11 (exponential smoothing with a constant of 0.4). The other moving average, Technique 4, produces the worst results of all techniques. The agreement between predicted and attained availabilities is poor for all techniques except Technique 11, and the second best, Technique 6, which is a variance-to-mean ratio of 3 on all items.

C.6 Results of the F-16 Demand Prediction Experiment

Table C-l 1 provides results for the 720 first-indenture items on the F-16 engine and airframe. The same 10 techniques were compared, and Technique 11 again came out best. It has a slightly lower attained availability than Technique 13, but the predicted availability is in better agreement. We note that many of the predicted availabilities are too high, because the flying hours and demands increased substantially over the 16 quarters. For example, the demand rate during the last 4 quarters was about 41% greater than during the first 12 quarters. However, excellent availability was attained with Technique 11, when this was not taken into account (except in the sense that exponential smoothing weights recent data more heavily in estimating the mean demand).

Some other observations concerning Table C-11 are noted here. The three Bayesian techniques (1, 2, and 4) and the Poisson technique (5) produced extremely low attained availabilities. Again Technique 11M (moving average) is far inferior to Technique 11 (exponential smoothing with a constant of 0.4).

In the F-16 case, we looked at the items that were primarily responsible for the degradation in each technique. The Bayesian techniques and the Poisson did badly on five or six items that showed a rapid increase in usage over the 12 quarter base period. The best variance-to-mean ratio performed much better on these items, particularly when it was used with the more responsive exponential smoothing instead of an 8 quarter moving average.

Because of the increase in the F-16 flying-hour program, we consider demand estimates based on demand per flying hour in the next section.

Table C-11. **Availability of F-16 Engine/Airframe: $80 million budget**

		Availability (%)					
	Cost	Attained	Predicted	Attained by Quarter			
Technique	($M)	Annual	Annual	1	2	3	4
1	120.04	2.2	100.0	8.7	0.0	0.1	
2	120.01	8.2	100.0	32.7	0.0	0.0	0.1
4	120.08	0.0	100.0	0.0	0.0	0.0	0.0
5	120.01	7.8	100.0	31.0	0.0	0.0	0.0
6	120.02	37.6	97.9	88.9	36.0	24.5	1.1
11M	120.01	40.7	99.0	83.5	34.3	35.8	9.4
11*	120.00	90.5*	98.0	95.1	86.7	95.7	84.4
13	120.02	90.6	74.9	94.5	86.3	94.3	87.4
15	120.00	79.0	69.5	94.8	79.0	81.4	60.9
17	120.07	39.5	59.0	56.0	43.8	34.0	24.3

* Best technique.

C.7 Demand Prediction for F-16 using Flying Hour Data

The items were divided into two groups using the first 12 quarters of historical data as follows: Group A. lower variance-to-mean ratio for demand per quarter; Group B. lower variance-to-mean ratio for demand per flying hour.[1] Then two demand prediction techniques were applied to each group separately and compared:

1) Technique 11 of Table C-6: exponential smoothing with a smoothing constant of 0.4 to predict the mean and a variance-to-mean ratio given by Equation C.3 ($\hat{V} = 1 + .14\hat{m}^{.5}$), but not larger than 20.

2) A procedure based on research by Stevens and Hill (1973) that is used currently by the Air Force Logistics Command (AFLC) in some applications. It consists of an 8 quarter moving average for the mean and a variance-to-mean ratio, V, given by Equation C.1, $\hat{V} = 1.132(\hat{m}T)^{.3407}$. $\hat{m}T$ is the predicted average pipeline and V is constrained to be no less than 1 and no more than 5.

We used two techniques, because Technique 11 had performed best in all of the previous comparisons, and the AFLC technique is actually used in some Air Force applications. As shown above in Table C-9, the variance-to-mean ratio for the AFLC technique of Equation C.1 varies with the repair time.

The details of how the sample was divided into two groups follow. First, the original sample of 1,214 items from the D041 data system for the F-16 engine and airframe was reduced by 281 items for which there was no demand during the first 12 quarters. In a real application it may be necessary to stock some of these items, but both prediction techniques estimate zero demand. Of the remaining 933 items, demand per flying hour could not be computed for 159; the flying hours were missing in the item data for one or more quarters in which demand was positive. For each of the remaining 774 items, the following equations were computed over n = 12 quarters of history:

[1] For some items the program element is sorties or operating hours. Since the program element on most items is flying hours, we will use that terminology for any program element.

$$\bar{d} = \sum_{i=1}^{n} d_i / n \tag{C.4}$$

$$V_1 = \frac{1}{\bar{d}(n-1)} \sum_{i=1}^{n} (d_i - \bar{d})^2 \tag{C.5}$$

$$\bar{\delta} = \sum_{i=1}^{n} d_i / \sum_{i=1}^{n} F_i \tag{C.6}$$

$$V_2 = \frac{1}{\bar{\delta}(n-1)} \sum_{i=1}^{n} F_i \left(\frac{d_i}{F_i} - \bar{\delta}\right)^2 \tag{C.7}$$

where d_i and F_i are the demand and flying hours in quarter i. We have used the symbol $\bar{d}$ for the mean demand per quarter and $\bar{\delta}$ for the mean demand per flying hour, since the numerical values will be very different. Equations C.5 and C.7 are the variance-to-mean ratios for demand per quarter and demand per flying hour respectively. Equation C.5 is Equation 2.2 with division by n - 1 to provide an unbiased estimate of variance and by the mean $\bar{d}$ to yield a variance-to-mean ratio. Similar divisions are used in Equation C.7, which is justified at length in Hodges (1985). When Equation C.5 is multiplied by $(n - 1)$, it is the Poisson index of dispersion of Equation 4.38 with $(n - 1)$ degrees of freedom.

Of the 774 items for which both Equations C.5 and C.7 could be computed, the latter was smaller for 511 items which were placed in Group B. The balance of the original 933 items were put in Group A. This is summarized below in Table C-12.

Table C-12. Estimators A and B for F-16

	Total Items		Best Items	
	Est. A	Est. B	Est. A	Est. B
Number of items	933	774	422	511
Average variance-to-mean ratio	66.024	43.683	44.181	34.323

The average variance-to-mean ratio over the 12 quarter base period for each group of items is shown at the bottom of Table C-12. Note that these variance-to-mean ratios are used only for separating the items between Group A and Group B, not for predicting demand. The reason for breaking the sample into two groups was to use demand per quarter or demand per flying hour, depending on stability.

Table C-13 shows that over the entire group of 933 items with some positive demand during the first 12 quarters, average demand per item increased from 22.0622 in quarter 1 to 50.9528 in quarter 16 (131% increase). When split into the two groups according to stability, average demand for Group A items increased from 21.3483 in quarter 1 to 30.3175 in quarter 16 (42% increase); the average demand per flying hour for Group B items increased from 27.0444/1,326 = 0.0204 in quarter 1 to 70.1919/2,154 = 0.0326 in quarter 16 (59% increase). The percentages for each of the two groups are much lower than the original percentage of 131% for all items, indicating that the program element does improve stability over time.

Table C-13. Average Demand/Item by Quarter for F-16

Quarter	All 933 Items, Demand/Item	Group A (422 Items), Demand/Item	Group B (511 Items)	
			Demand/Item	Fly Hours/It
1	22.0622	21.3483	27.0444	1.326
2	25.0054	22.8839	31.6505	1,430
3	29.1629	23.8294	39.7060	1,567
4	27.9378	22.3389	38.4273	1,538
5	29.2765	23.7749	39.0995	1,552
6	30.4373	21.3152	43.7991	1,667
7	31.9625	20.5190	47.6622	1,850
8	35.7728	21.8175	54.4347	1,817
9	39.7717	27.0142	52.4633	1,756
10	45.0772	28.6422	60.6680	1,868
11	47.8617	29.2062	65.3131	1,969
12	46.8810	27.4360	64.9737	1,951
13	47.3655	29.1043	64.4646	2,011
14	49.5531	30.6682	67.2545	2,149
15	51.2776	30.1872	70.9152	2,242
16	50.9528	30.3175	70.1919	2,154
% Increase Qtr. 16/Qtr. 1	131	42	159	62

We are making the assumption that an item placed in Group B because of a lower variance-to-mean ratio in the 12 quarter base period can be predicted more accurately from demands/flying hour. Of course, this really depends on how accurately we can forecast flying hours over the next year. If we had no such ability, a forecast of demand/flying hour would be useless. Earlier research has shown that if our flying hour forecasts are within about 20% of the true value, demand/flying hour will give better predictions than demand/quarter for items in Group B. Since planned flying constitutes a major portion of the flying hour program, a forecast within 20% seems likely.

The final number of items was reduced to a multiple of 20 for ease of processing, i.e., 420 items for Group A and 500 items for Group B. A budget of $110 million was divided in rough proportion to the number in each group, $50 million and $60 million, respectively.

Table C-14 shows the results for items in Group A and Table C-15 the results for Group B. The procedure was to use the first 12 quarters of D041 to predict the mean and variance-to-mean ratio over the next 4 quarters, as described above. Then an optimal availability model was used to allocate investment across items. Actual demands during the next 4 quarters were used in a simulation to assess prediction accuracy as in the earlier comparisons. The attained availability was computed across the group of items on a daily basis and averaged over each quarter. Both attained availability under no cannibalization and with cannibalization are shown.

Technique 11 performs better than the AFLC technique in all comparisons. The availabilities with no cannibalization are more relevant, because that was the objective we set out to maximize with our availability model. Thus it is not surprising that the availability differences under cannibalization are smaller.

Table C-14. **Availability (%) Group A: 420 Items Demand per Quarter: $50M Budget**

	No Cannibalization		Cannibalization	
Quarter	Technique 11	AFLC Technique	Technique 11	AFLC Technique
1	83.7	63.7	90.9	81.6
2	67.8	45.0	84.3	78.8
3	80.7	58.2	90.4	80.0
4	62.2	45.9	80.2	79.5
Average	74.6	53.2	86.5	80.0
Backorders	28.7	61.1		

However, a comparison of availabilities tends to underestimate the margin of superiority enjoyed by Technique 11. For example, in Table C-14 an improvement in availability from 53.2% to 74.6% is more than a 40% improvement, because availability does not increase linearly with cost. (Figure 2.5 showed a typical availability versus cost curve where each successive increase in availability is more expensive). A better estimate of the margin of superiority for Technique 11 is the average number of backorders, which are reduced by 53% in Table C-14 and 57% in Table C-15.

The overall availability for the 920 items of the F-16 is obtained by multiplying the availabilities for group A and B. The availabilities are 43.3% for Technique 11 and 20.1% for the AFLC technique. Since these availabilities are so low and to insure that the superiority of Technique 11 does not depend on the budget level, we performed the analysis again with a budget of $180 million split equally between groups A and B. This resulted in availabilities of 70.4% for Technique 11 and 40.8% for the AFLC technique; backorders averaged three times as large under the latter technique.

Table C-15. **Availability (%) Group B: 500 Items Demand per Flying Hour: $60M Budget**

	No Cannibalization		Cannibalization	
Quarter	Tech. 11	AFLC Tech.	Tech. 11	AFLC Tech.
1	93.6	84.1	95.8	89.9
2	87.8	58.2	94.9	30.2
3	27.1	6.6	32.9	16.5
4	23.9	2.3	31.1	13.3
Average	58.1	37.8	63.7	50.0
Backorders	50.2	117.9		

One technical note should be added. The variance-to-mean ratios over the first 12 quarters (3 years) exceed 44 for demand (Group A) and 34 for demand per flying hour (Group B), as can be seen from the bottom of Table C-11. These variance-to-mean ratios are extremely high, because they are computed over a long period of time, using the 12 quarter means; in fact, we believe that the means are changing, which would reduce the variance, but in an unknown way. By contrast, the variance-to-mean ratios in our predictions from Technique 11 are much smaller (exceeding 10 for only 4 items and never reaching the limit of 20). The reason is that the variance-to-mean ratio in the predictions is only over a year, and it increases with the length of the period.

When the backorders on each item are examined, it appears that about two-thirds of the total improvement due to Technique 11 is because of exponential smoothing, one-third to the variance-to-mean ratio formula and the higher limit of 20.

Similar calculations were performed for the A-10 with comparable results. For example, on the 320 items where demand per program element was the better predictor the average availability with no cannibalization and a budget of $50 million was 94.3% for Technique 11 and 85.5% for the AFLC technique. The average backorders were 5.9 and 19.6, respectively.

C.8 Correlations

In this section we want to look at the "physics" of the demand process by examining the correlation of demand in different quarters. The correlation coefficient is defined as:

$$\rho = \frac{\sum (x_i y_i - \overline{xy})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} \tag{C.8}$$

where x_i and y_i are the observations to be correlated, $\bar{x}$ and $\bar{y}$ are the means. The correlation coefficient is defined between -1 and +1. In many applications the x_i and y_i are observations from two different sets of data. Here we want to compute the autocorrelation function where the y_i are identical with the x_i but displaced in time. For example, if each $y_i = x_i + 3$, the function is called the autocorrelation with lag 3. Since there are 16 quarters of data on each item, there are 15 pairs of adjoining quarters (lag 1), 14 that are separated by a quarter (lag 2), etc. Our objective is to examine the behavior of the autocorrelation function as the lag increases from 1 to 15.

Unfortunately, the data series on each item is very short. As the lag increases, the number of pairs available for autocorrelation decreases. With 16 quarters we can only compute meaningful correlations for the first 11 lags or so. Each computed correlation can be compared with the tabled value in a statistics text to test[1] whether it is significantly different from zero. For example, with the 15 observations at lag 1 the observed correlation must

[1] The tables are percentiles of the distribution of observed correlations when $\rho = 0$.

exceed .514 to be significant at the 95% level; with the 5 observations at lag 11 the observed correlation must exceed .878.

With a rapidly increasing flying-hour program, such as that of the F-16, it is appropriate to compute the correlations between demand per program element since the program element change may differ from item to item. When this is done for all 615 items on the F-16 with program element information in each of the 16 quarters, the results are shown in Table C-16.

Over the entire group of 615 items, the average correlation is significant at the 95% level for intervals of lengths 1 through 3. On an individual item basis, it is seen that many items have a significant positive correlation for intervals of lengths 1 through 4. After that point (a 1 year separation), the number of positive correlations that are significant tends to be only slightly larger than the number of significant negative correlations. Note that with a 95% significance level and 615 items, we would expect that about 31 items would appear to have significant correlations, positive or negative, by chance.

Table C-16. **Correlations between Demand per Program Element: 615 F-16 Items**

Number of Quarters Apart	By Item: Average Correlation	By Item: Number Significant +	By Item: Number Significant −	Across-Items Correlation
1	.201	151	11	.936
2	.085	91	18	.903
3	.049	225	20	.883
4	.029	64	26	.888
5	−.026	39	29	.883
6	−.030	40	33	.869
7	−.034	25	28	.854
8	.022	53	27	.849
9	.013	43	24	.845
10	−.019	33	28	.829
11	−.038	25	23	.807
12				.810
13				.826
14				.825
15				.749

In order to examine the relationship between demands in quarters with greater lags, we have used the correlation formula from Equation C.8 in a second, non-standard way. Instead of $\bar{x}$ and $\bar{y}$ being the means over the quarters for an item, we let them be the means over all items and quarters. This increases the sample size dramatically, so that even at lag 15 the number of pairs equals the the number of items. These "correlations" are shown in the last column of Table C-16. Note that the values are near one because each (x_i, y_i) pair is from a particular item, but the means, $\bar{x}$ and $\bar{y}$, are from all items. The numerical values themselves are meaningless since they will be larger if the item means have greater dispersion, and vice-versa.

We cannot use statistical tests of significance on these values in the last column of Table C-16, but we can look at how they change as a function of the number of quarters between observations. Although the values do not decrease uniformly, there is a clearly decreasing pattern that confirms the correlation results by item. The implication of this analysis is that demand prediction techniques should give more weight to more recent data. Exponential smoothing is such a technique, and this helps to explain why its performance was so much better than the moving average.

Table C-17 shows that analysis of the A-10 aircraft yields similar results. over the entire group of 444 items, the average correlation is significant at the 95% level for intervals of lengths 1 through 4; the number of items with significant positive correlation is also substantially greater than the number with significant negative correlation for intervals 1 through 4.

Table C-17. **Correlations of Demand per Program Element: 444 A-10 Items**

Number of Quarters Apart	By Item			Across-Items Correlation
	Average Correlation	Number Significant		
		Positive Corr.	Negative Corr.	
1	.207	117	3	.921
2	.095	70	13	.910
3	.064	83	5	.897
4	.059	48	17	.886
5	.001	39	22	.869
6	−.048	19	26	.872
7	−.022	16	17	.864
8	.011	22	13	.849
9	.011	22	11	.851
10	−.038	18	16	.844
11	−.056	9	18	.850
12				.847
13				.857
14				.862
15				.822

All data analyzed in this report have been quarterly D041 data, where all base demands on a given aircraft type have been aggregated. It is important to be sure that demand at an individual base is not fundamentally different from the sum of demand at all bases; that demand over shorter time periods does not behave differently. For this reason, we obtained the transaction data showing the date of each demand at a base during a 2 year period. The base was England Air Force Base, Mississippi, with A-10s, and the data covered the period from April 1982 through March 1984. The demands for each repairable item were aggregated in 2-week periods as the fundamental unit of analysis. The smallest meaningful period would have been a week, since we do not want the lighter flying schedule on weekends to bias the results. A 2 week period was selected so that demands during the period would be slightly larger; this gave us 52 periods for correlational analysis.

The results of similar correlational analysis on the 2-week demands at England Air Force Base are shown in Table C-18. Some major differences are noted from the previous correlations. The number of items whose correlations are positive and statistically significant when the periods are adjoining is only 95/741, or 12.8%, as compared with about 25% for Tables C.15 and C.16. The reason is that the average demands per period are much

Table C-18. Correlations of Demand per 2-week Period: 741 A-10 Items

	By Item			
		Number Significant		
Number of Quarters Apart	Average Correlation	+	−	Across-Items Correlation
1	.034	95	6	.505
2	.027	76	1	.478
3	.021	81	5	.465
4	.013	59	5	.443
5	.000	43	5	.445
6	.006	45	1	.449
7	−.003	39	3	.427
8	−.004	33	0	.427
9	−.007	26	2	.413
10	−.015	18	3	.407
11	−.021	22	6	.395
12	−.015	27	6	.387
13	−.007	36	7	.384
14	−.017	29	2	.384
15	−.015	22	10	.389
16	−.024	20	9	.364
17	−.020	29	8	.373
18	−/024	17	8	.360
19	−.024	12	8	.346
20	−.020	25	10	.341

smaller and the variability in this number across items is much smaller (the average item demand per 2 week period at England Air Force Base is about 0.2, as compared with quarterly demand across all F-16 bases of 35).

The correlations in the last column of Table C-18 decrease fairly uniformly (only the first 20 intervals are shown in the table, but the correlations decrease to .201 for the largest separation of 51 periods). The first value of .505 is much smaller than the first value in the other correlational analyses across items in Tables C.15 and C.16, due to the low values of demand. However, the qualitative behavior of the demand for 2-week periods is consistent with the D041 data providing further support for our major conclusions.

C.9 Smaller Smoothing Constant for Low-Demand Items

There was some suspicion that predictions for very low-demand items could be improved by incorporation of a longer history, equivalent to a smaller constant in the exponential smoothing. Our procedure was to use a smoothing constant of 0.1 on any items with fewer than 10 demands in the first 12 quarters (typically about 20% of the items). Though the predicted availabilities were very slightly higher, the actual attained availabilities were slightly higher in only one case and lower in seven. This is shown in Table C-19. It should be noted that the values of 0.1 and 10 demands are arbitrary. Different values will give somewhat different results.

In Table C-19, the attained availabilities are generally higher than the predicted availabilities for the C-5 engine, lower for the A-10 and F-16. The reason is that the annual dollar value of demand for the C-5 engine decreased during the period being predicted to only 62.6% of its value in the base period, while the value increased 19.5% for the A-10 and increased 41.4% for the F-16.

However, the allocation of investment appears to have been very good. Results for an 8-quarter moving average are shown at the highest budget level for each weapon system. Note that in two of the three cases, the moving average leads to severe degradation in attained availability.

In summary, Technique 11 with a smoothing constant of 0.4 for all items appears to be best.

Table C-19. Availabilities With Different Smoothing Constants

Budget ($M)	Smoothing Constant: .4 for All Items — Attained	Predicted	.1 for Items with Fewer Than 10 Demands — Attained	Predicted
		C-5 Engine		
10	79.6	61.4	79.6	61.5
15	87.8	80.3	87.8	80.5
20	95.5	91.7	95.6	91.7
25	97.7	97.2	97.4	97.2
25	97.8[1]	97.5[1]		
		A-10 Airframe		
40	30.9	20.5	30.3	20.7
60	84.7	78.8	84.4	78.9
80	96.2	97.2	96.0	97.2
100	98.7	99.7	98.7	99.7
100	60.1[1]	99.9[1]		
		F-16 Engine/Airframe		
60	22.5	35.9	22.4	36.1
90	73.3	80.8	72.9	80.9
120	92.7	96.2	92.5	96.3
150	97.4	99.4	97.4	99.4
150	39.2[1]	99.8[1]		

[1] Indicates the result of an 8-quarter moving average used to predict the mean rather than exponential smoothing.

C.10 Summary

The major conclusions are summarized below:

1. Demand prediction techniques should give more weight to recent data, because the mean demand tends to drift with time. Demand in any period of time tends to be more highly correlated with demand in adjoining time periods than with those further away.

2. Exponential smoothing is consistently the best technique of those tested for estimating mean demand. With quarterly data, a smoothing constant of about 0.4 appears best. (or about 0.15 on monthly data).

3. Use of Poisson or a constant variance-to-mean ratio assumption leads to poor allocations of investment, because system availability attained during the prediction period is low. For each aircraft type analyzed (C-5, A-10, F-16), a "good" choice for variance-to-mean ratio, *V,* over a predicted year as a function of the estimated annual mean, *m,* is:

$$\hat{V} = 1 + .14\hat{m}^{.5} \tag{C.9}$$

with a maximum value of 20. For a fixed investment of $80 million on the C-5 airframe, the Poisson assumption leads to an attained availability of 56.6%; the formula above leads to 82.8% for the next year. Alternatively, the same availability could be attained with 20% less investment and the variance-to-mean ratio formula above.

Our previous work, including Slay and Sherbrooke (1988), indicates that a Poisson process is still the best description of how demands are generated. But, the assumption of a Poisson process with stationary increments where the mean stays constant is the most optimistic and unrealistic type of Poisson process.

Note that the independent variable, *m,* is the annual mean, rather than the average pipeline as used by AFLC. It is the drift in mean demand over the prediction period that is important. Equation C.9 above for the variance-to-mean ratio in combination with exponential smoothing for mean demand led to backorder reductions of over 50% for the F-16 and A-10 when compared to the AFLC variance-to-mean formula combined with an 8 quarter moving average.

This system approach to estimating variance-to-mean ratio is much better than using past history on each item. The latter is too unstable, particularly when the mean demand is changing.

4. When program element data, such as flying hours, are available and demand per program element is more stable, they should be used for prediction. If demand per quarter is more stable over the base period, that should be used, instead. In our analyses, slightly more than half the items were in the former group. The same prediction technique is recommended for both groups of items, but the choice of technique is more critical for the former group.

5. The use of an availability model to evaluate the accuracy of prediction techniques is a powerful approach, and is highly recommended to other investigators. We noted that previous research by others such as Stevens and Hill (1973) has provided estimating relationships for the variance-to-mean ratio, but failed to evaluate the predictive accuracy.

The findings below are less definitive than the conclusions above, because they depend on the techniques used in this study. It is possible that other researchers may devise new techniques that will alter one or more of these findings:

1. Techniques for predicting trend were unsuccessful. We believe that the lack of success with second-order smoothing, observed in our earlier research as well, is due to a phenomenon called "regression to the mean".[1] Since mean demand drifts with time, this implies that variance increases as the period being predicted moves into the future or becomes longer.
2. Combinations of techniques did not improve predictions. An example of a combination technique is a variance-to-mean ratio that is the average of a power curve prediction and a sample estimate based on historical data for the item.
3. Bayesian techniques used in this study were good but not best. The attained availabilities were fairly high in most cases, but the predicted availabilities were almost always too high. The good performance was due in part to some arbitrary adjustments that had to be made for high demand items. Bayes is more complicated and has the further limitation that it cannot be used for initial procurements, since it uses demand data.
4. A smaller exponential smoothing constant for low-demand items did not improve predictions.

These conclusions and findings appear to hold consistently across the C-5, A-10, and F-16 weapon systems as represented in 16 quarters of D041 data. Further support was found in an analysis of daily demand data over a 2 year period from a single A-10 base, England Air Force Base. Data has been analyzed, both with and without program element information.

Some words of caution are in order. There are a few items with very large and highly variable demand. System availability is critically influenced by how well the prediction technique performs on those items. Unfortunately, there is no way around this fact of life, which seems to be true of most complex systems. This is not a defect in the evaluation measure; rather, it is related to the "physics" of our problem. The practical implications are that the technique for the variance-to-mean ratio that is "best" for one weapon system is not likely to be "best" for all others,

[1] There is a story about pilot training in the Israeli Air Force some years ago. The instructors noticed that after a particularly good landing causing praise from the instructors, the next landing tended to be closer to an average quality landing; after a poor landing and some "chewing out", the next landing tended to be better. This is simply an example of "regression to the mean" where a particularly good or bad landing is more likely to be followed by one closer to average quality, regardless of instructor feedback.

principally because of the behavior of a few items. However, we have found techniques that appear to do a good job across weapon systems and appear to be substantially better than techniques now in use.

We have done some preliminary work on demand prediction over periods of time longer than a year or further in the future as reported in Sherbrooke (1987). The results suggest that the variance-to-mean ratio in these cases should be larger. However, there is a practical question as to whether all of the variance in demand should be reflected in the inventory policies, as the repair system has an important adaptive capability to change repair priorities.

Another point of caution is that a number of items show no demand over the base period. Having no basis for picking one prediction technique over another for these items, we have excluded them from predictions. There are two implications: First, it is desirable to retain data history longer to improve detection of items whose "true" mean demand is not zero. Second, it may be necessary to estimate some positive demand for items with no observed demand (e.g., using a Bayesian procedure as discussed in Section 4.10). If this is not done, the resulting value for aircraft availability is likely to be unsatisfactory.

We have attempted to understand the "physics" of the demand process, not just compare different demand prediction techniques. It is our hope that the results of our analysis will be useful in other applications, and that the approach itself will have even greater applicability. While no claim to optimality can be made for our empirical work on demand prediction, it is important to reiterate that the optimal availability models in this book are useful only if the required data can be estimated.

Appendix F shows a Windows tool, the Demand Analysis System, that can be used to perform all the analyses above in an efficient fashion.

Appendix D

PREDICTING WARTIME DEMAND FOR AIRCRAFT SPARES

If an elderly but distinguished scientist says that something is possible he is almost certainly right, but if he says it is impossible he is very probably wrong.
-Arthur C. Clarke

D.1 Appendix Overview

The research reported on here was performed by the author in collaboration with Michael Slay of the Logistics Management Institute and Lt. Colonel Dave Peterson of the Air Force (Slay et al. 1996). There is also a more detailed report (Sherbrooke 1997). The research came about because of the surprisingly small number of demands for spares in Desert Storm during 1991, despite the large number of flying hours. Longer sorties were flown than in peacetime, but the demand per sortie remained about the same.

Since peacetime data are used to forecast wartime requirements and peacetime sorties are much shorter than those of Desert Storm and probable future wartime scenarios, it is critical to know whether spares demand is driven by sorties, flying hours, or some combination of them. In this appendix we review the Desert Storm experience, review the literature, consider the advantages of a controlled experiment, and then report on our analysis of over 200,000 sorties. Although this appendix is concerned with the military and wartime planning, it has implications for commercial aircraft and peacetime spares prediction as well.

D.2 Desert Storm Experience

Table D-1. Desert Storm Spares Demand

		F-15C/D	F-16C/D
% of Planned	30 day flying hours	236%	142%
	30 day sorties	85%	91%
Item demand rates	Over-predicted>25%	84%	81%
	Within +/- 25%	7%	10%
	Under-predicted>25%	9%	9%
# items best predicted by	Flying hours	58	23
	Sorties	214	117

Table D-1 shows the results for one F-15C/D squadron of 24 aircraft during the first 30 days of Desert Storm; they flew 236% of the planned flying hours, but only 85% of the planned sorties. Actual demand was much less than forecast on the basis of flying hours. Item-by-item predictions based on sorties were better for 214 items as compared with 58 for flying hours. Similar results were obtained for 72 F-16C/D aircraft.

D.3 Literature Review

Table D-2 summarizes some Air Force studies on the effect of sortie duration on spares demand. A slope of 100% (45 degrees on a graph where sortie length is plotted on the horizontal axis and spares demand on the vertical axis) indicates that spares demand is proportional to sortie length, whereas a slope of 0% indicates that sortie duration has no effect.

Table D-2. Regressions of Maintenance Removals on Sortie Duration

Aircraft	Systems	% Slope	# Sorties	Author	Date
C-5A	All	5	79,181	Shaw	1981
C-5A	Engine	8	79,181	Shaw	1981
C-141	All	28	835,000	Shaw	1981
C-141	All	22	73,000	Shaw	1981
C-141	All	100	50,388	Howell	1978
C-130E	All	33	45,000	Shaw, Howell	1981, 1978
727	All	-1	3,300	Shaw	1981
P-3C	All	77	54,892	Howell	1978
B-52D	All	20	10,809	Boeing	1970

Most of the slopes are more consistent with the assumption that sortie length has little or no effect (the slope for the P-3C is negative, but not significantly different from zero). There are only two exceptions: 1) a slope of 77% for the 727; 2) a slope of 100% for the C-141 in a 1978 study. The first is of limited relevance because the 727, a commercial aircraft, flew very short sorties of 45 minutes to 1½ hours, and the demand rates are an order of magnitude less than for military aircraft; the second is an anomaly because the 50,000 sorties were drawn in an unknown way from the 835,000 sorties used in the 1980 study of the C-141 which resulted in a slope of only 28%.

The results are of limited relevance to our problem of tactical aircraft flying multi-hour sorties. The studies are all over 20 years out of date, and the studies were primarily briefings without detailed methodological support.

D.4 Proposal for a Controlled Experiment

We suggested to the Air Force that a controlled experiment would be desirable to study the effects of sortie duration on the demand for spares on tactical aircraft. The experiment would control for age and material condition of the aircraft, pilot proficiency, and other factors that might contaminate an uncontrolled study.

As an example of the problems that can arise in an experiment without control, consider the hypothetical situation depicted in Table D-3. Assume that there is some new maintenance/operating procedure that is applied to half the fleet chosen at random. Since the number of broken aircraft (one or more demands per sortie) in the control group exceeds the number in the treatment group, we would conclude that the new treatment is not effective.

Table D-3. Random assignment of aircraft to treatment and control groups

Number of aircraft	Number Broken	Total	% Broken
Treatment	28	50	56
Control	22	50	44

But, suppose some bright-eyed analyst then comes up and says, “We should have broken the aircraft into two groups based on age, because the older aircraft are likely to require more spares.” After the results are broken down by age, they might look like Table D-4. Note that in each of the two age-specific groups the treatment aircraft now look better than the control aircraft, contradicting our earlier results. (Of course, the totals in the sub-groups must add up to the group totals in Table D-3, e.g. the broken aircraft in the two age-specific treatment groups of 26 + 2 = 28 of the combined

Table D-3). This is a classical statistical problem known as Simpson's Paradox.

It is difficult to know what to conclude. The problem arises, of course, because more of the older aircraft happened to land in the treatment group in our hypothetical experiment. This problem would not have arisen if the original experiment had controlled for aircraft age. Similarly, we wanted the Air Force to have a controlled experiment with controls on any variable suspected to affect spares demand. If we controlled on variables that later had no affect, it would make no difference to the analysis.

Table D-4. Data of Table D-3 Broken into Older and Newer Aircraft Groups

Older Aircraft	Number Broken	Total	% Broken
Treatment	26	40	65
Control	8	10	80
Newer Aircraft			
Treatment	2	10	20
Control	14	40	35

Unfortunately, the Air Force decided against a controlled experiment on the grounds that it would severely impact their peacetime training program.

D.5 Data Analysis – F-15 C/D Aircraft

The data source was the Core Automated Maintenance System (CAMS) which supplied: 1) operational data such as tail number, sortie length, time of the sortie, takeoff and landing locations, and mission type; 2) maintenance data such as tail number, start time, work unit code (item identification), how malfunctioned, when discovered, and action taken. The supply data could not be related to sorties, so we had to use the maintenance data as a proxy for supply data. We used only that maintenance data that reflected a remove-and-replace action, excluding time-change items and aircraft modifications.

For the 68 aircraft based at Langley, Va. We found that the highest number of demands occurred on the last sortie of the day, as shown in Table D-5. The number of demands tended to decrease by sortie when there were multiple sorties, until that last sortie. This is an important variable that had never been identified in earlier studies, but tactical aircraft with training sorties of 1-2 hours have the opportunity of flying many more sorties in a day than the aircraft in Table D-2 (one aircraft in Table D-5 flew 7 sorties). The sortie lengths preceding the last sortie of the day also tended to be shorter as the number of sorties by an aircraft increased, but far less dramatically.

A logical question is whether the last sortie of the day is the result of a grounding failure, or planning. It is clear from several sources of evidence that the last sortie of the day is usually planned, and that the higher maintenance following the last (or only) sortie is the result of deferred maintenance. One example of the deferred maintenance is that the average demand per sortie originating and landing at Langley, the home base, is 0.34. On sorties originating at Langley, but landing elsewhere the rate is 0.19, whereas sorties originating elsewhere and terminating at Langley had a rate of 0.67.

Table D-5. Impact of Sortie Number on Langley F-15C/D Demand

Sortie Number of the Day	Number of Sorties	Average Length (Hours)	Average Demands/Sortie
Only Sortie	1,857	1.54	0.62
1 of Multiple	2,804	1.35	0.17
2 of Multiple	796	1.22	0.14
3 of Multiple	418	1.15	0.12
4 of Multiple	178	1.12	0.10
5 of Multiple	45	1.00	0.11
6 of Multiple	1	0.90	0.00
Last of Multiple	2,820	1.33	0.52
Overall Total/ Weighted Average	8,919	1.36	0.37

The type of mission also had a large effect on demand per sortie. In aerial combat training, the aircraft may pull as much as eight G's. Longer missions such as cross-country and training deployment are less stressful as seen in Table D-6. If we did not control for mission type, the effect of sortie length on demand would be understated.

Table D-6. Impact of Mission Type on F-15C/D Demand

Mission Type	Number of Sorties	Average Length (Hours)	Average Demands/Sortie
Aerial Combat Training	7,247	1.32	0.39
Cross Country	498	1.47	0.15
Training Deployment	973	1.64	0.27
Other	201	1.23	0.56
Overall Total / Weighted Average	8,919	1.36	0.37

We analyzed the group of 7,108 aerial combat training missions that took off from and landed at Langley. One of the highest demand rates was for the 88 sorties lasting less than 0.8 hours. We excluded these because of the high prevalence of functional check flights and air aborts. Functional test flights are very short, very high-failure post-maintenance test flights (where the failures are really related to an earlier flight). We would have retained the air aborts if we had known the planned sortie durations, but these were not collected.

The resulting regression was for 7,020 sorties with durations between 0.8 and 7.3 hours. The regression slope was 18% which is statistically significant at the 95% level (there is less than a 5% chance that such a large slope could occur by chance if there were no relationship between sortie duration and demand).

When we add another independent variable for last (or only) sortie of the day, the slope of demand as a function of sortie length drops to 13% (still statistically significant). The reduction in slope occurs because the last (or only) sortie of the day, which has more demand, tends to be slightly longer as shown in Table D-5.

D.6 Analysis of Other Data Sets

We analyzed several other data sets as well with the results shown in Table D-7. The tactical aircraft are listed first, followed by the A-10 used in close air support, two bombers (B-1 and B-52H), a transport (C-130H), a gunship (AC-130) and a tanker (KC-135). The F-16 C/D had the largest slope for tactical aircraft, after correcting for the influence of last sortie, tied at 13% with the Langley F-15 C/D analyzed above. The largest corrected slope for all aircraft was 20% for the B-52H, which was the same slope that was estimated for the B-52D in Table D-2; the latter was probably the best

Table D-7. Slope % of Demand vs. Sortie Length by Aircraft Type

Aircraft Type	Before Sortie #	After Sortie #	Number of Sorties
F-15C/D Langley	18*	13*	7,020
F-15C/D 1993 All	0	0	16,522
F-15C/D 1994 All	0	0	15,514
F-15A	0	0	13,287
F-15E	0	0	11,942
F-16C/D	21*	13*	61,499
F-111F	0	0	3,401
F-117A	0	0	8,794
A-10, OA-10	10*	6*	30,967
B-1	2*	8*	7,754
B-52H	13*	20*	3,197
C-130H	6*	3*	27,919
AC-130	12*	11*	1,247
KC-135	9*	8*	17,504
Overall	10*	7*	226,565

*** indicates significance at the 95% level**

earlier study because it contained roughly equal numbers of 4, 8, and 12 hour sorties from three bases in Southeast Asia, providing a particularly good basis to estimate the slope of demand versus sortie length.

One interesting result came about in the analysis of the A-10, OA-10 aircraft. The original regression showed a very steep slope, unlike any other data set. When we examined the data more closely, we found that the bases with higher average demand (and slightly longer sorties) were regular USAF bases; the others were Reserve and National Guard bases. This is shown in Figure D-1, where each data point on the graph represents the mid-point for a group of sorties at a particular location. When we analyzed the two groups separately and combined the results, we obtained the slopes shown in Table D-7 for the A-10, OA-10.

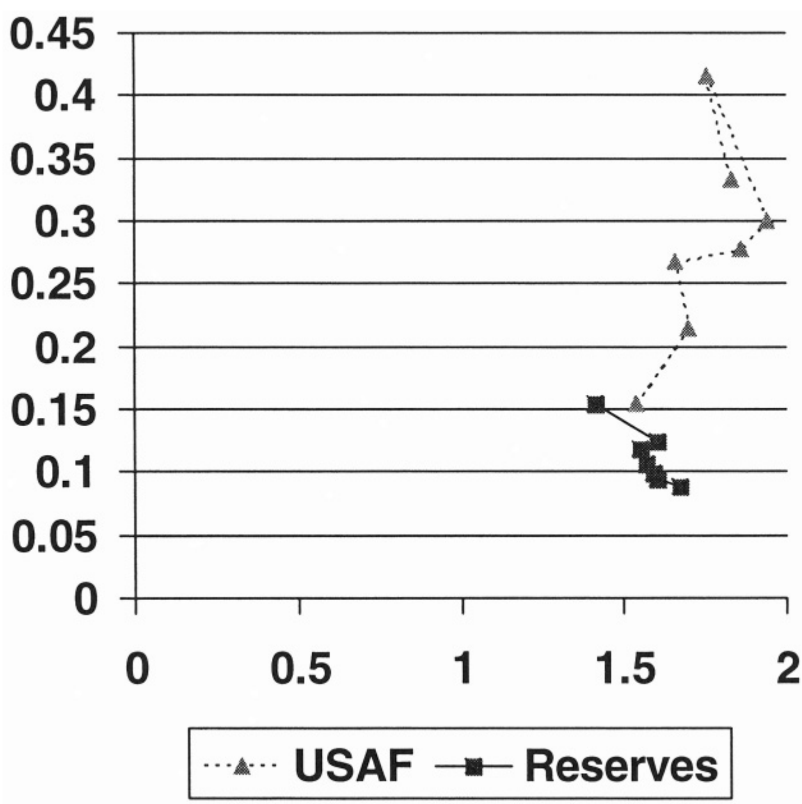


Figure D-1. **Demands (y axis) vs. Sortie Length (x axis) for A-10 aircraft**

We don't know exactly why the USAF bases had such high demand rates. It has been suggested that the Reserve and National Guard pilots are usually retired USAF and thus have more experience; perhaps more importantly the same is true for maintenance personnel. Another factor is that some of the Reserve and National Guard bases are located near metropolitan areas (e.g. Kansas City), where the use of live ordnance is restricted, also contributing to the shorter sortie lengths. For the purposes of the analysis of demand versus sortie length, the explanation is not critical, provided that we analyze the two groups separately.

D.7 Summary

We asked ourselves if there could have been any reasons in our analysis that a real effect of sortie length influencing demand might have been obscured. Some of the possibilities include:

1). There were very few long sorties in most of the data sets, making it difficult to estimate the slope. For example, in the case of the F-15C/D, 50% of the sorties were between 1.1 and 1.5 hours and 90% were between 0.4 and 2.1 hours.

2) Some of the shorter sorties may be more stressful missions, even within a group such as aerial combat. For example, the F-15C/D can pull 8 G's with a center-line fuel tank, but only 6 G's with wing tanks. The latter tend to be longer sorties, and probably have fewer demands; but, the mission code is the same in the CAMS data.

3) Demands are estimated from maintenance actions, rather than supply data. Unfortunately, we had no alternative.

Nevertheless, the results of Table D-7 suggest that sortie length has a modest effect on demand. The slopes for all the data sets average 10% before and 7% after adjustment for the last sortie effect, and both are statistically significant. Only one slope exceeded 13% (the B-52H at 20%).

Some of our other conclusions were:

1) Demand rates increase slightly with the number of days since the last sortie (F15C/D, F-16C/D), but the results were not statistically significant.

2) Aircraft do not consistently have higher or lower demand rates. That is, there appears to be no potential for assignment bias of "peaches" or "lemons" to particular missions (F-16C/D).

3) We did find larger slopes of 30% for Electronic Warfare Systems and 20% for Fire Control (F-16C/D). However, this may be partly due to the fact that those systems tend to get exercised only on longer missions (see above discussion of the A-10 results for the USAF versus Reserve and National Guard).

The Air Force implemented our conclusions, settling on a 10% slope for the effect of sortie duration on spares demand for tactical aircraft. They still collect peacetime demand data on a per-flying-hour basis. The demand per flying hour, d, is converted for wartime planning, so that a tactical aircraft sortie of x hours is assumed to have $d + (x - 1)d/10$ demands – a 10% slope for sortie duration.

The Air Force estimates that it would have overstated wartime spare requirements for tactical fighters by $1.1 billion in 1993 otherwise. We are inclined to quote the late Senator Dirksen, "A billion here, a billion there. Pretty soon it adds up to real money."

Appendix E

VMETRIC MODEL IMPLEMENTATION

Common sense is the collection of prejuduces acquired by age eighteen.
- Albert Einstein

E.1 Chapter Overview

In this chapter we illustrate an implementation of the VARI-METRIC theory in the VMetric model. The purpose is to demonstrate that personal computers can provide sufficient power and elegant user interfaces. It is not intended as a user's manual or a tutorial on how to build the most efficient computer program. The VMetric model is a product of the Tools for Decision (TFD) group, located in Monterey, CA., whose web-site is www.tfdg.com. The VMetric engine was written by the author.

The several Windows computer screens are shown as a user would encounter them in the process of doing a VMetric optimization problem. Of course, the limitations of black-and-white reductions required to fit this book do not do justice to the actual appearance of the screens.

We list the variables used by the model. Only a small number of variables on each item are actually *required,* but with hundreds or thousands of items in a typical computation the data management tasks are significant;

there are a number of *optional* variables that allow the user to fine-tune his solution to particular needs.

VMetric has no limitation on the number of items, sites, indentures, or echelons. It does require that the sites can be organized into a tree structure as shown in Figure 1-3, where each site has a specific support site identified, either within the organization or a contractor outside. The physical location of the support sites for different items may differ. (For example, the Air Force has several depots, each of which handles certain types of items).

We will use a standard application to illustrate VMetric, but the reader is reminded of the large number of different applications that can be optimized with VMetric as discussed in Chapter 9.

E.2 VMetric Screens

When VMetric begins, Figure E-1 appears. The logic is that after loading a file, the user normally uses the buttons from left to right as described below. In this illustration we have loaded an actual Coast Guard data base with 951 items on the HH-60 Helicopter.

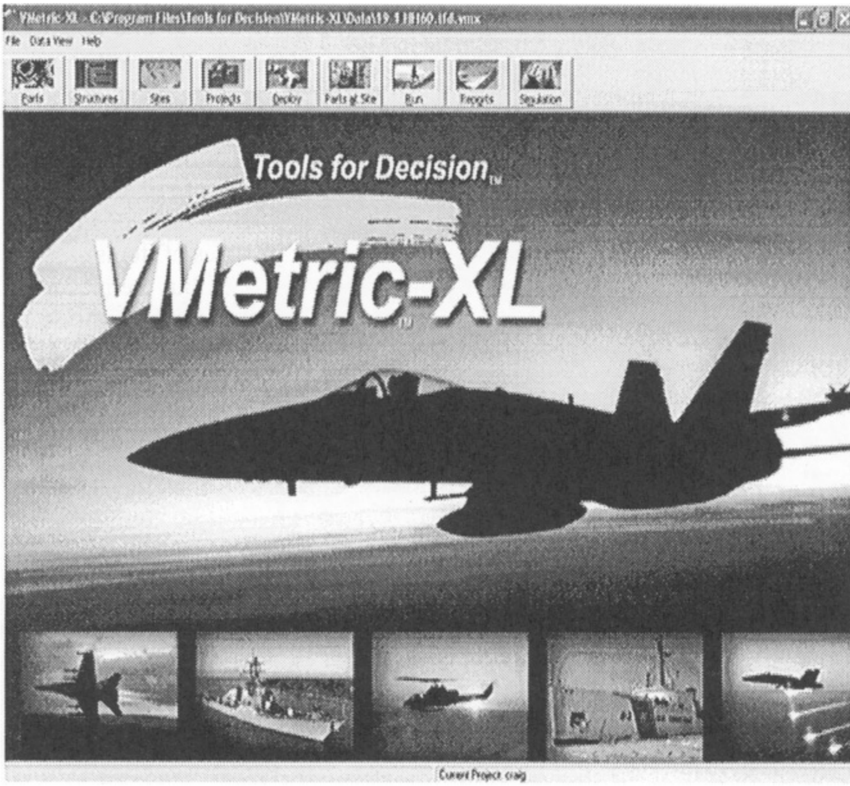


Figure E-1. **VMetric Welcome Screen**

When the user clicks on the left-most button, the Parts Library appears as in Figure E-2. A number of variables appear in the column headings, some of which are required and some of which are optional (the columns for required information are colored, but it is impossible to distinguish in black and white). Part uniqueness is determined by the combination of reference number or, if no reference number, item name and Commercial and Government Entity (CAGE) Code with a null CAGE being allowed. So only item name or reference number must be populated to generate a unique part. All other identifying fields are optional, including National Stock Number (NSN):

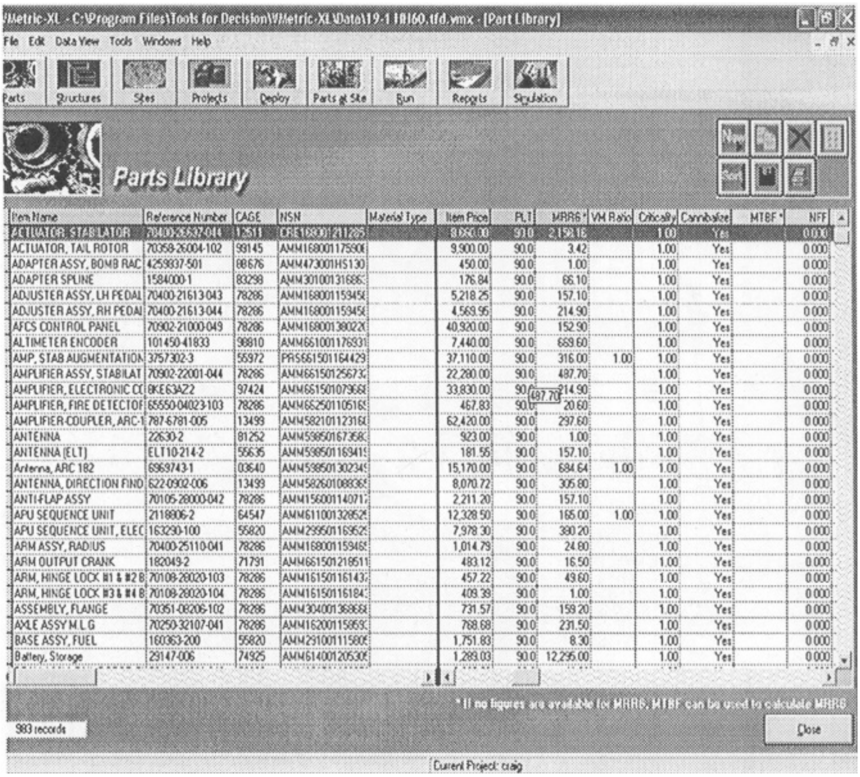


Figure E-2. VMetric Parts Library

Required Variables

- Item Name
- Item Price
- PLT - Procurement Lead Time
- MRR6 - Demands per Million Hours (alternatively MTBF, NFF, and RIP below can be entered).

Optional Variables

- Reference Number
- CAGE code - Manufacturer identification
- National Stock Number

Material Class - When left blank the default is miscellaneous (Misc). It is possible to enter other material classes such as electronics or airframe. Then in the default screen below, one set of values for Repair Cycle Time, Order and Ship Time, and NRTS can be entered and automatically copied for all items in a material class. Very useful in early provisioning.
VM Ratio - Item override of the global variance-to-mean ratio from the power curve $V = 1 + Am^B$ where m is average annual demand.
Criticality - Default criticality (of LRUs) is 1; can be set higher or lower to reflect relative importance of a backorder.
MTBF - Mean Time Between Failures (instead of MRR6)
NFF - No Fault Found Rate in the range 0-1 (default = 0)
RIP - Repair in Place in the range 0-1 (default = 0)
Resale Price - If resale of existing stock is an option
Setup Cost - If one or more units of an item is purchased in an initial procurement, this cost is added

When the user clicks on the second button from the left, Structures, the indenture structure of items from the Parts Library of Figure E-2 appears. The level of indenting in Figure E-3 indicates the indenture. For example, the Engine is an LRU, followed by two SRUs, the second of which contains two sub-SRUs (indenture level 3). This convention gives an unambiguous definition of all the parent-child relationships. The parent of any item is that item above it in the list which is closest, similarly for the grandparent. Thus, the LINER, COMBUSTION is a child of Module, Hot Section, and a grand-child of ENGINE.

In this view there are two additional variables by item, as well as two that can be accessed with the configuration tab:

Optional Variables
Weight - Only used in multiple constraint problems
Volume - Only used in multiple constraint problems

Required Variables
QPA(n) - Quantity per next higher assembly
QPA(k) - Quantity *required* for next higher assembly to operate. Only for LRUs where redundancy is allowed in the case of a single end-item.

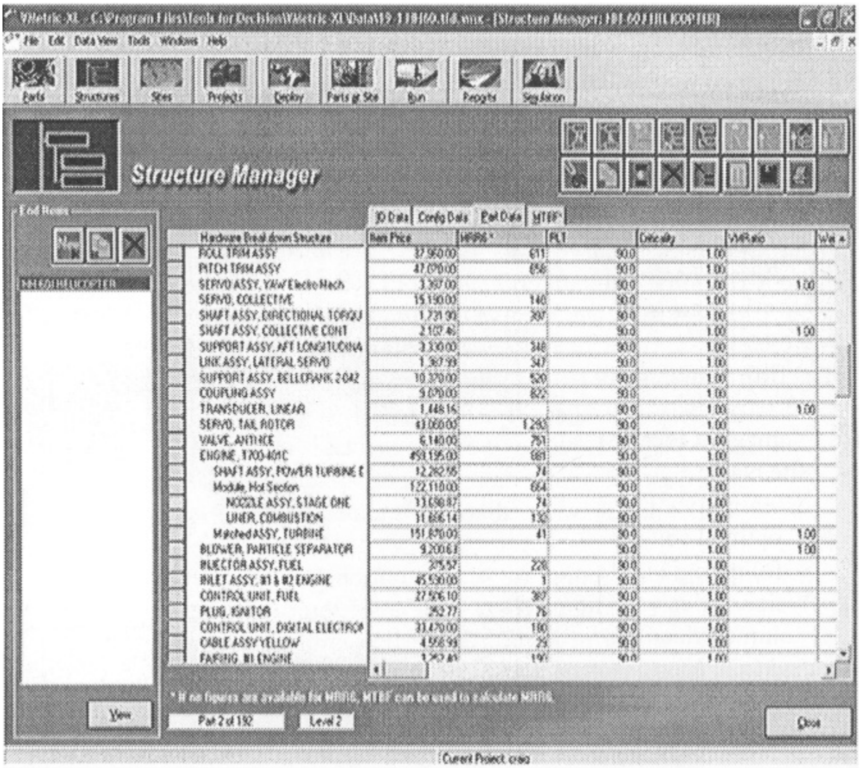


***Figure E-3.* VMetric Structure Manager**

The third button, Site Library, is just a list of site names with their support site. In general there will be one site whose support site is contractor for the replenishment of condemnations. The fourth button, Projects, enables the user to give the run an identifying name and description as well as some additional system variables. The first five variables are used to calculate reorder points and order quantities; they are optional because they do not affect initial procurement quantities.

Optional Variables

- Internal Order Cost - The cost of ordering from a support site within the logistics system. Usually a nominal value like $5.
- External Order Cost - The cost of placing an order on a contractor.
- Order Cost Threshold - An item cost times quantity above which the high order cost is used for external orders
- Order Cost High

Holding cost rate - An annual rate that reflects the cost of borrowing, warehouse costs, and some obsolescence.
Lateral Supply Time - The number of days that it takes to ship a depot-repairable LRU from a base with stock to a base with a backorder.
Variance-to-mean ratio parameters, A, B, and VMax for the power curve relationship: $V = 1 + Am^B$ where m is the average annual demand. The default is Poisson with a constant mean.
Resale Factor- If resale price is not set, resale factor is multiplied by item price to generate a resale price if resale is allowed.

The fifth button, Deploy, shows the following by site:

Number of identical sites of this type (default is 1)
Site Name
Essentiality - These are weights applied to the sites to obtain an optimal solution for site specific goals (see the footnote following Figure E-7).
Weight Price (Optional - only used for multiple constraints)[1]
Volume Price (Optional - only used for multiple constraints)
End Item Name

When the user clicks on the End Item Name, the details are provided for each site:

Number of the end-item type at the site
Name of the end-item
Operating Hours/week
Maintenance Availability (Optional - default is 1)

There are a set of tabs on this screen, one for each material type. Misc. is the default. (We have a second material type in the illustration called joe). This provides information on the following for each operating site:

Support Site
Repair Cycle Time
Order and Shipping Time
NRTS

[1] The weight price and volume price are Lagrange multipliers which must be adjusted iteratively to meet the site weight and volume constraints. They are precisely analogous to the site essentialities discussed in the footnote following Figure E-7. See also Figure 6-3.

When the Misc. tab is selected, the above information is displayed as shown in Figure E-4. The display also shows three sites including Clearwater, all of which are supported by Clearwater Support (as discussed in Section 9.9). Similarly, the display shows that all other sites including AR&SC Ops are supported by AR&SC. This is an example of what we mean by ragged echelons as depicted in Figure 1-3.

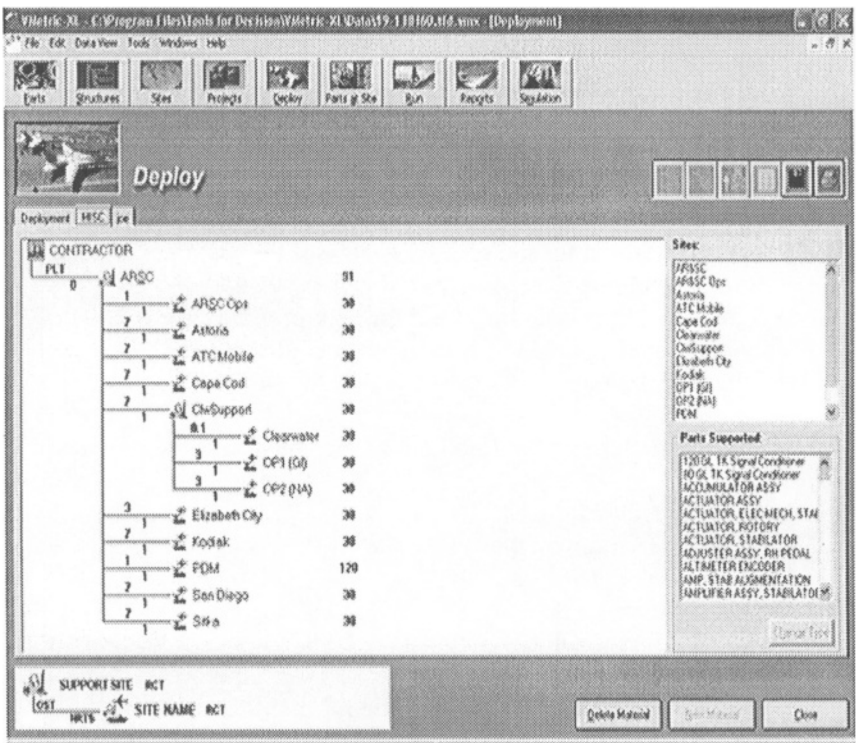


Figure E-4. **VMetric Deployment**

As mentioned earlier, the idea of the material types is that if one is assigned to each item, it is possible to set defaults easily for the latter three quantities on all items of a material type. This is particularly useful in early model runs when the data is still tentative.

The sixth button is Parts at Site, a complete tableau of every variable at every site. The user need not enter anything on this screen as it has been populated by the earlier screens; however, the user may override any entry, in which case the entry is shown in bold-face to distinguish it from default values entered elsewhere.

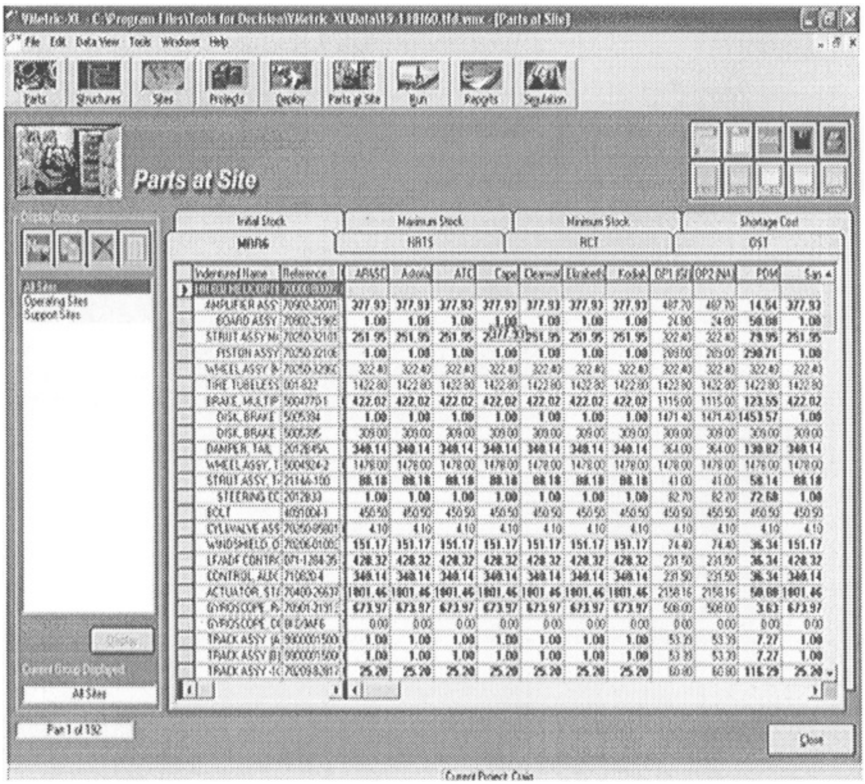


Figure E-5. **VMetric Parts at Site**

We are now ready to make a VMetric model run. The seventh button provides many choices in two groups, Starting and Stopping:

Starting Choices

Fill pipelines to a specified fraction (shown to the right) - Used primarily for preliminary runs to speed up the calculation.

Use Initial Stock - Yes indicates that stockage should include these.

Use Minimum Stock - Yes indicates min stocks by location for each item should be used.

Include Initial Stock in Budget - Yes indicates that the value of initial stocks should be added to the cost of items bought.

Use Resale Price - Yes indicates that Sale Prices should be used, if present, or Resale Factor to allow owned items to be sold if optimal.
Use Shadow Prices - Yes means that the problem is multi-resource (e.g. weight or volume as well as cost). Use the weight and volume prices by site.
Use deterministic lot size calculation - If the economic order quantity at the depot for reprocurement times the item cost exceeds the order cost threshold, use the larger external order cost
Redistribute Initial Stock - Yes means that the stock already owned should be optimally redistributed before more is purchased.
Use Maximum Stock - Yes indicates max stocks by location and overall for each item should be used.[1]
Use Shortage Cost - Yes indicates that values by item and location should be used.to weight backorders. Default is 1.
Use Setup Cost - Yes indicates that this cost should be added to the first unit of an item that is procured.
Emulate MOD-METRIC - Yes causes VMetric to ignore demand for SRUs and below that generates during repair at support sites (see Section 5.4).

Stopping Choices

Evaluate Only - Yes causes VMetric to evaluate initial stocks only.
Set Goals by Project, by Site, by System, or by System and Site.
Goals - One or more of the following with the goal value:

Operational Availability
Budget Constraint
Slope **A_0%/Cost(000)**
Fill Rate
Average Days Delay

[1] The maximum stock capability was used in the applications of Sections 9.4, 9.9, and 9.13. It is important not to use maximum stocks on too many items, particularly with a high availability goal, as it may prevent VMetric from achieving the goal.

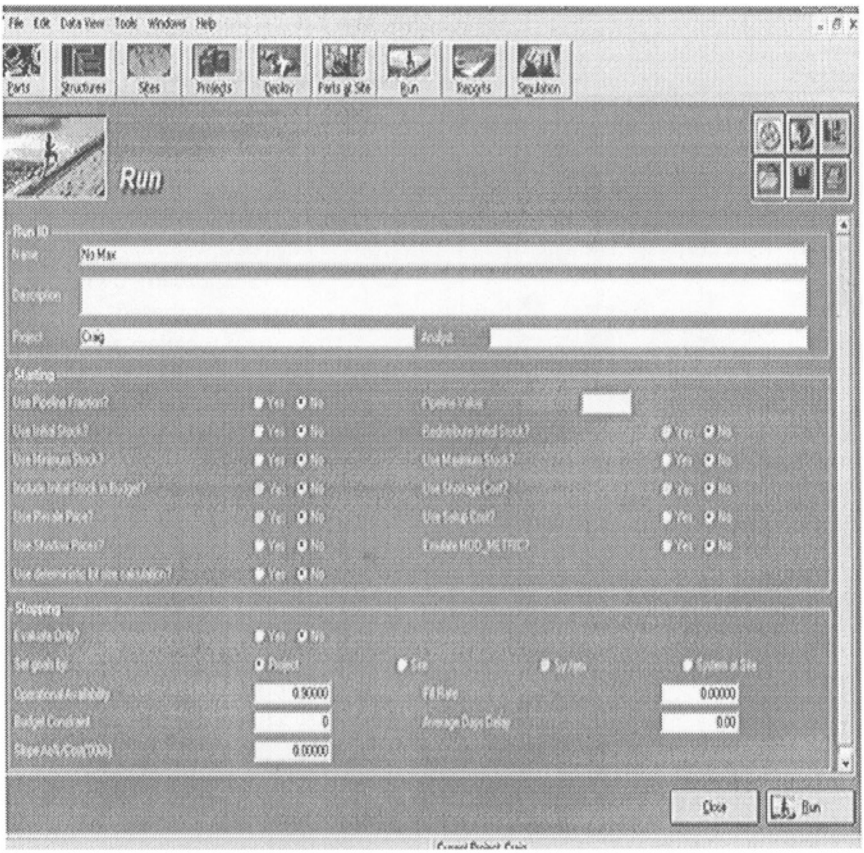


Figure E-6. **VMetric Run Screen**

When we make the run, an availability fraction of 0.90094 is obtained for the Project at a cost of $60,507,180. The individual site availabilities range from 0.84080 to 0.94530. Let's see how much more investment is required to obtain .90 goals by site using the Goals by Site button instead of the Goals by Project in Figure E-6.

Now the Progress Screen looks like Figure E-7 with the overall availability vs. cost screen on the right and the individual sites on the left with meters showing their availability at each step of the calculation. When the calculation is finished. Figure E-8 appears. The summary screen shows a budget of $65,897,160 (8.9% more than for the .90 Project availability goal) and an overall availability of 0.92257. The individual site availabilities can be seen by using the Summary by Site tab which shows a range from 0.90200 to 0.93889.[1]

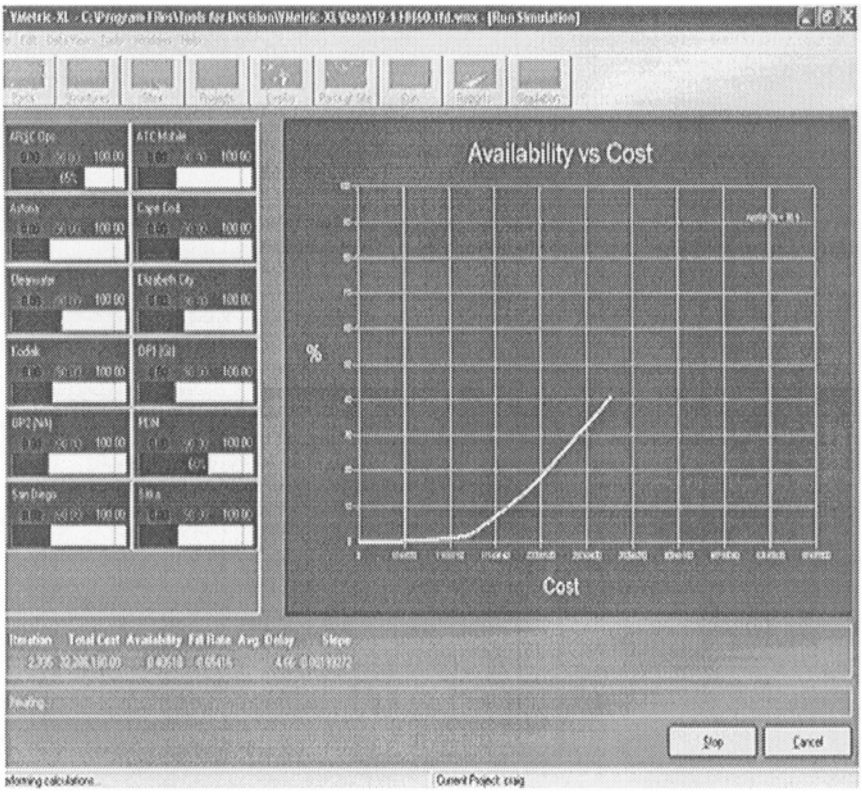


Figure E-7. **Availability vs. Cost Progress Screen**

[1] If the site essentialities are adjusted by trial and error (this took two iterations), it is possible to meet the individual site availability goals of .90 in a more optimal fashion. It is easy to see that the correct site essentialities provide an optimal solution, because they are actually Lagrange multipliers that represent the cost of meeting each site constraint. The final essentialities ranged from 0.5 to 2.0 and the availabilities ranged from 0.902 to 0.924 by site. The overall availability was 0.910 and the budget of $62,805,270 was only 3.8% more than for the original Project goal availability of 0.90. See also Section 8.14.

When I was in graduate school, one of my mathematics instructors said that Lagrange multipliers for constraints made no intuitive sense to him. But to any economist (and most mathematicians), they are the prices of constrained resources.

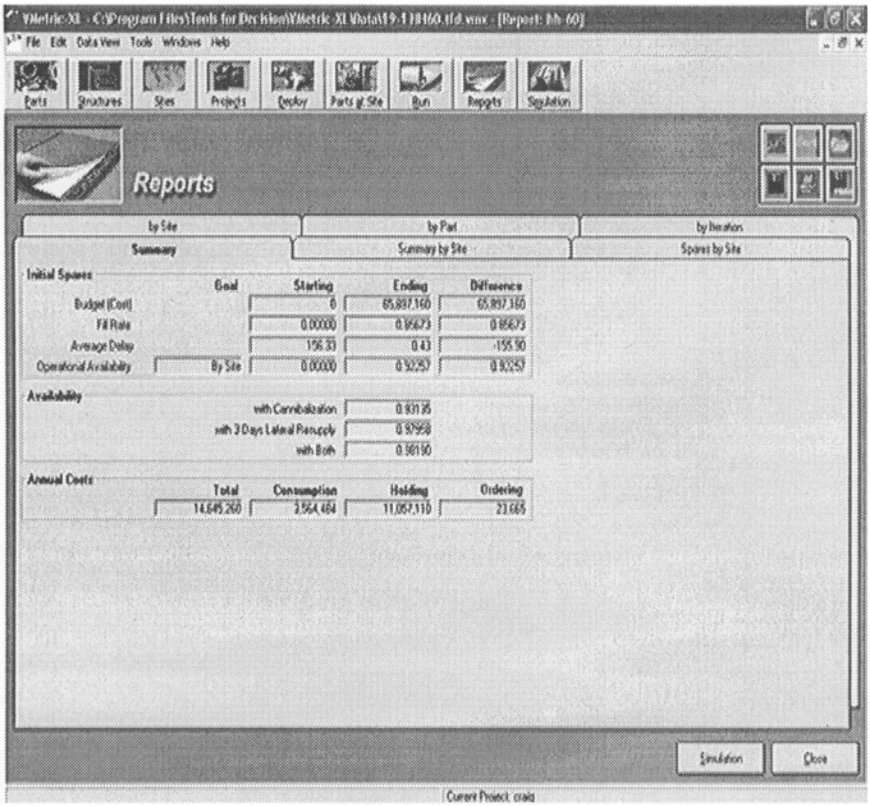


Figure E-8. **VMetric Output Report Screen for .90 Site Availabilities**

The Reports Screen also shows a Fill Rate of 85.6% for LRUs at operating sites and an Average Delay per LRU demand at operating sites of 0.43 days; the Availability with Cannibalization is 0.930135, Availability with a 3 Day Lateral Supply Time is 0.97958, and the Availability with Both is 0.98190. The reason that cannibalization shows such a small improvement is that there are very few helicopters at each site.

Annual costs of consumption or condemnations are $3,564,484. The annual holding and ordering costs are shown, but these are not out-of-pocket costs; they are accounting costs which have already been incurred by having a logistics system with item managers, equipment specialists, storage facilities, etc.

Of course, complete details are available by using the tabs: Summary by Site, Spares by Site, or Details by Site, Details by Item, and Details by Iteration. The latter shows the "shopping list" at each step of the calculation. Various graphical outputs are available as well.

Appendix F

DEMAND ANALYSIS SYSTEM

Part of the inhumanity of the computer is that, once it is completely programmed and working smoothly, it is completely honest. -Isaac Asimov

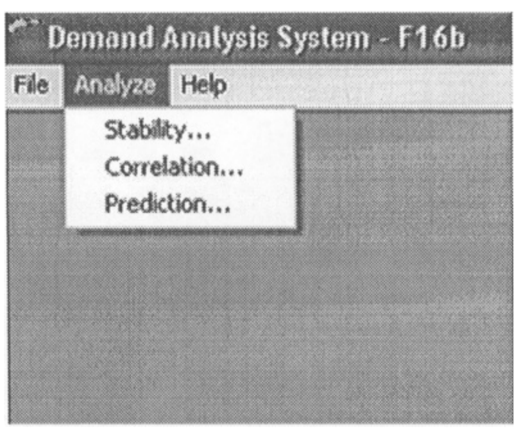


Figure F-1. **Types of Analysis in DAS**

The demand prediction studies in Appendix C were performed with a primitive version of the Demand Analysis System,[1] described in this appendix. The methodology was explained and justified in Section 4.15 and Appendix C so we will not repeat those arguments here.

First, we bring in a data file[2] with item data on unit cost, average repair time, and by period (monthly or quarterly) demand, flying hours, sorties, number of end-items. Then there are three types of analyses that can be performed as shown in Figure F-1: stability, correlation, and prediction.

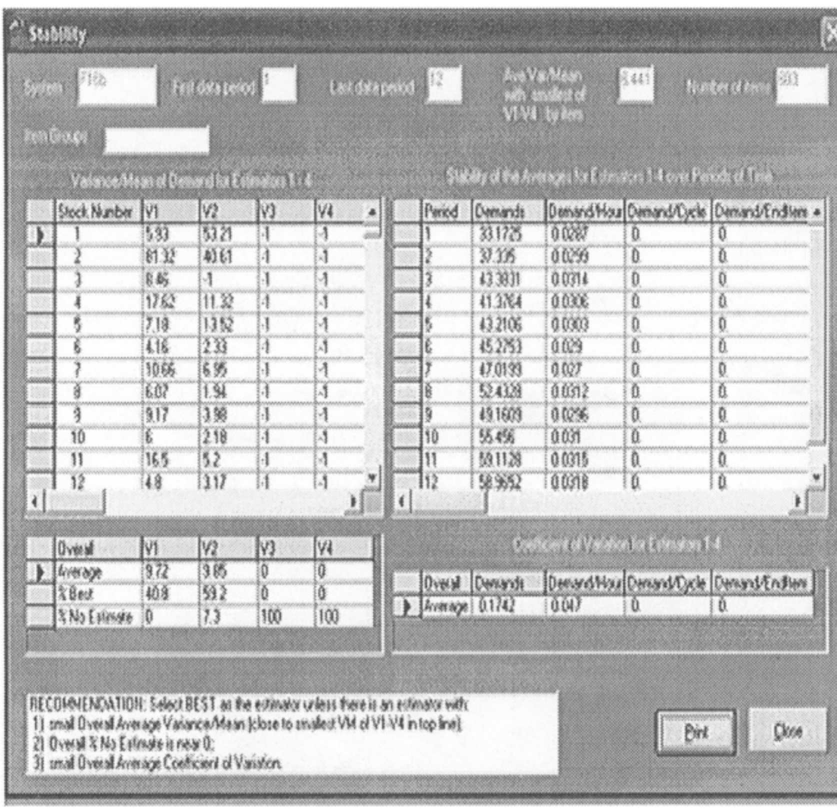


Figure F-2. **DAS Stability Analysis**

[1] DAS has a Windows interface with brilliant colors that the reader will have to imagine.

[2] This is data on the F-16, not the b version, and these 605 items are not the same as the 933 in Appendix C.

Stability analysis has the ability to compare up to four quantities such as demand, demand/flying hour, demand/sortie, and demand/end-item to see which is most stable over the base period (12 quarters in the cases of Appendix C). In Figure F-2 there were data only for the first two quantities on the F-16. On the right-hand side of Figure F-2 demands have a coefficient of variation (square root of demand divided by the mean) of 0.174 as compared to demand/hour whose coefficient of variation is 0.047. Clearly demand per hour is much more stable, and should be selected (the number of aircraft and flying hours increased dramatically over the 12 months). On the left hand side of Figure F-2 are the individual item variance-to-mean ratios computed over the 12 quarters. Some are very high (one exceeds 81), suggesting that the means over 12 quarters are not really constant.

Then correlation analysis is used to determine whether exponential smoothing is recommended. It can be seen from Figure F-3 that 9.784% of items have a positive correlation that is significant at the 95% level; since this is substantially higher than the threshold of 2.5% which would be appear to be significant by chance, exponential smoothing looks attractive. DAS computes an estimate of 0.167 and recommends exponential smoothing.

Figure F-4 compares the predictive accuracy of three procedures for the mean and variance-to-mean ratio: (1) an 8 quarter moving average and Poisson variance-to-mean ratio = 1; (2) exponential smoothing with a constant of 0.2 and Poisson variance-to-mean ratio = 1; (3) exponential smoothing with a constant of 0.2 and variance-to-mean ratio:

$$\hat{V} = 1 + .14\hat{m}^{.5}$$

with a maximum of 20.

The base period is quarters 1-12 and the prediction period is quarters 13-16. A total of \$145M is allocated by each procedure. Note that we have selected demand /hour because of the earlier stability analysis.

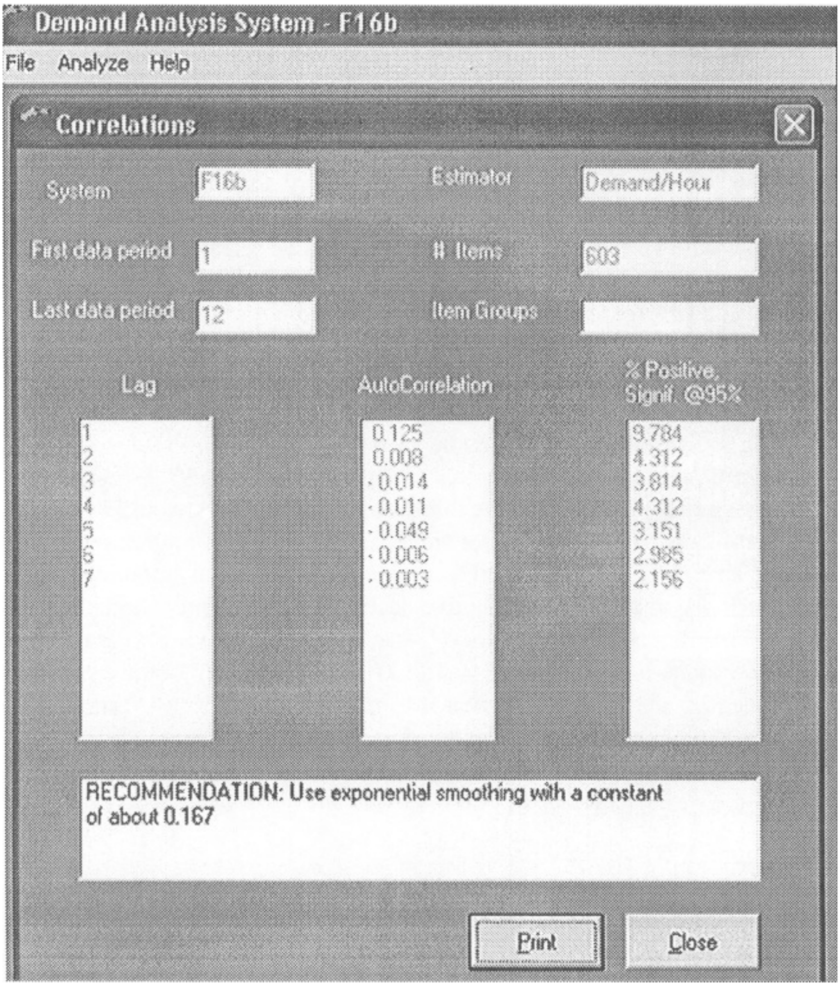


Figure F-3. **Autocorrelations for various lags**

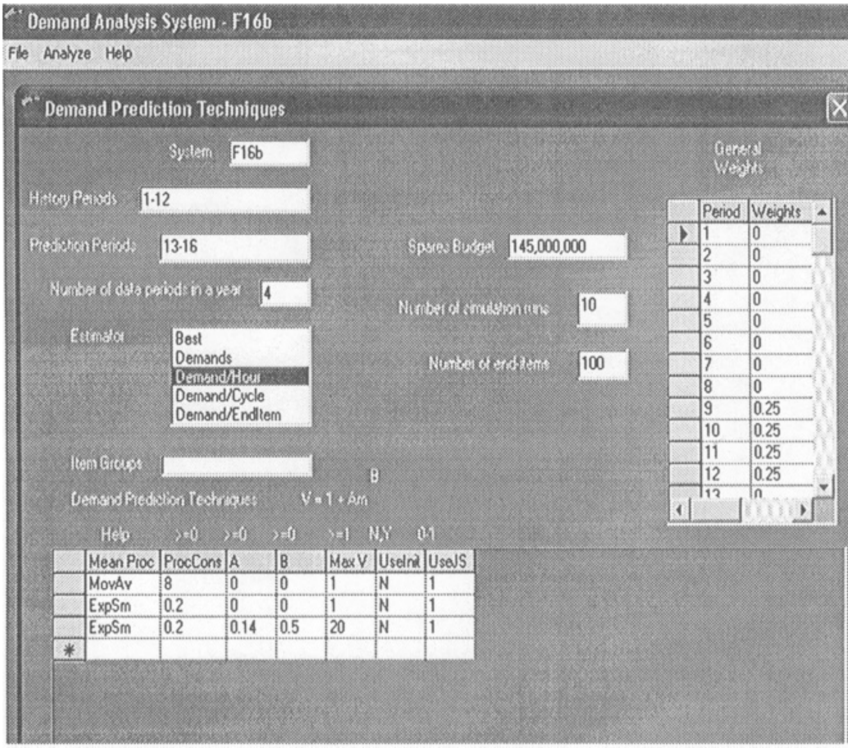


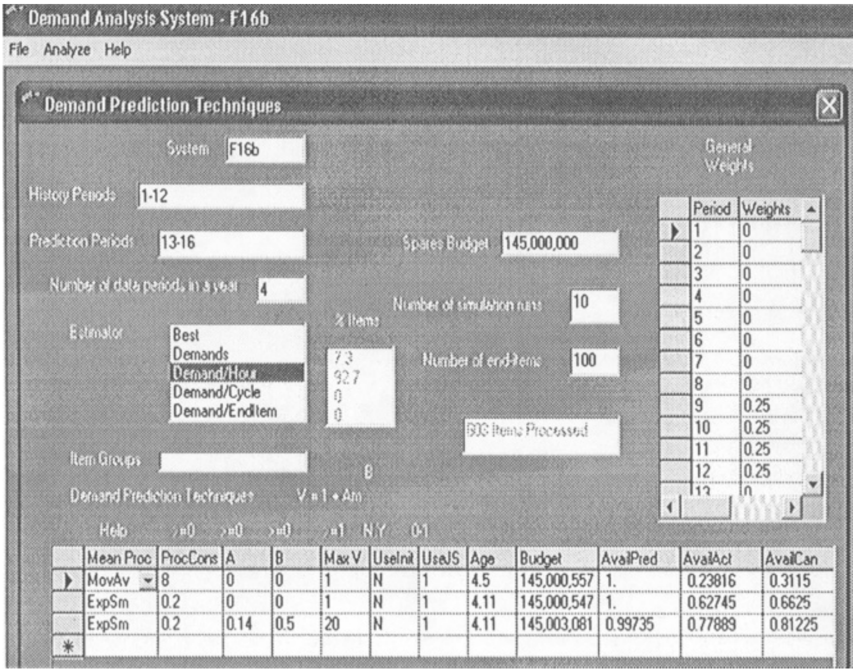
Figure F-4. **Comparison of 3 Procedures**

In Figure F-5 we see the results of the three procedures. While the predicted availabilities were all near one, the "actual" availabilities over the prediction year were 24%, 63%, and 78% for no cannibalization and 31%, 66% and 81% with cannibalization.

The quarterly availabilities shown in Figure F-6 are even more revealing. Looking at the "actual" availability columns, we see that procedure 3 has the lowest availability in the first predicted quarter of the three procedures; but it wins in the other three quarters and is incredibly high at 85% in the last predicted quarter, suggesting that it might perform well even in later quarters, because it takes into account means that change with time.

In conclusion we ask the reader to think of how much more budget he would normally require to move from an availability of 24% to 78%. Here it is obtained at no extra cost by proper demand prediction procedures.

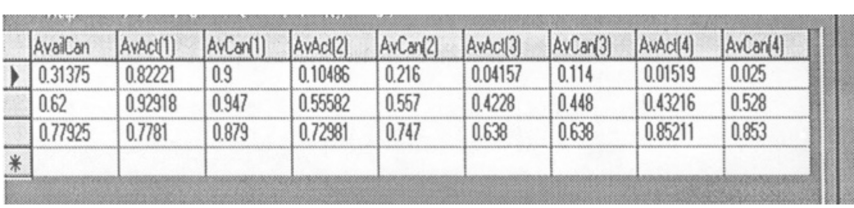
DAS automatically loads VMetric with the optimal demand rates by item obtained with the best demand prediction procedure. Normally a user would rerun DAS to update the item demand rates prior to making a VMetric optimization.



Period	Weights
1	0
2	0
3	0
4	0
5	0
6	0
7	0
8	0
9	0.25
10	0.25
11	0.25
12	0.25

Mean Proc	ProcCons	A	B	Max V	UseInit	UseJS	Age	Budget	AvailPred	AvailAct	AvailCan
MovAv	8	0	0	1	N	1	4.5	145,000,557	1.	0.23816	0.3115
ExpSm	0.2	0	0	1	N	1	4.11	145,000,547	1.	0.62745	0.6625
ExpSm	0.2	0.14	0.5	20	N	1	4.11	145,003,081	0.99735	0.77889	0.81225

Figure F-5. **Results of Comparing 3 Procedures**



AvailCan	AvAct(1)	AvCan(1)	AvAct(2)	AvCan(2)	AvAct(3)	AvCan(3)	AvAct(4)	AvCan(4)
0.31375	0.82221	0.9	0.10486	0.216	0.04157	0.114	0.01519	0.025
0.62	0.92918	0.947	0.55582	0.557	0.4228	0.448	0.43216	0.528
0.77925	0.7781	0.879	0.72981	0.747	0.638	0.638	0.85211	0.853

Figure F-6. **Quarterly Details for 3 Predictions**

REFERENCES

Abell, J. B., L. W. Miller, C. E. Neumann, J. E. Payne (1992). *DRIVE (Distribution and Repair in Variable Environments).* RAND Corporation, R-3888-AF, Santa Monica, CA.

Air Force Logistics Command (1967). *Base Stockage Model Report.* Wright-Patterson AFB, OH. 9 May.

Air Force Material Command (1999). *A Comparison of Alternative Business Rules for Repair and Distribution Prioritization,* XPS, Wright-Patterson AFB, OH. 3 May.

Axsater, S. (1990). Modelling Emergency Lateral Transshipments in Inventory Systems. *Management Science* 36, 1329-1338.

Boeing Military Airplane Systems Division (1970). *B-52 Operations in Southeast Asia versus CONUS.* D162-10015-1, Seattle, WA.

Clark, A. and H. Scarf (1960). Optimal Policies for a Multi-Echelon Inventory Problem. *Management Science* 6, 475-490.

Clark, A. (1981). Experiences with a Multi-Indentured, Multi-Echelon Inventory Model. In Schwarz, L. B. (ed), *Multi-Level Production/Inventory Control Systems: Theory and Practice.* North-Holland, Amsterdam, pp. 229-330.

Cohen, M., P. V. Kamesam, P. Kleindorfer, H. Lee, and A. Tekerian (1990). Optimizer: IBM's Multi-Echelon Inventory System for Managing Service Logistics. *Interfaces.* 20, 65-82.

Crawford G. B. (1981). *Palm's Theorem for Nonstationary Processes.* RAND Corporation, R-2750-RC, Santa Monica, CA.

Efron, B. and C. Morris (1977). Stein's Paradox in Statistics. *Scientific American,* 119-127. Vol 236:18, May.

Ehrhardt, R. (1979). The Power Approximation for Computing (*s*,*S*) Inventory Policies. *Management Science* 25, 777-786.

Federgruen, A. and P. Zipkin (1984). Computational Issues in an Infinite-Horizon, Multi-echelon Inventory Model. *Operations Research* 32, 818-836.

Feeney, G. J. and C.C. Sherbrooke (1965). *An Objective Bayes Approach for Inventory Decisions.* RAND Corporation, RM-4362-PR, Santa Monica, CA.

Feeney, G. J. and C.C. Sherbrooke (1966). The (s-1 ,s) Inventory Policy Under Compound Poisson Demand. *Management Science* 12, 391-411.

Feller, W. (1958). *An Introduction to Probability Theory and Its Applications. Vol I.* John Wiley & Sons, New York.

Graves, S. C. (1985). A Multi-Echelon Inventory Model for a Repairable Item with One-for-One Replenishment. *Management Science* 31, 1247-1256.

Groover, C. W., S. J. Culosi, L. L. Woods, et. al. (1987). The Logistics Information System Analysis (LISA). Vol. 2. *Analysis of Stock Control and Distribution System,* Systems Research and Applications Corp., Arlington, VA.

Gross, D., and C. M. Harris (1974). *Fundamentals of Queueing Theory.* John Wiley & Sons, New York, 415-419.

Gross, D. (1982). On the Ample Service Assumptions of Palm's Theorem in Inventory Modeling. *Management Science* 28, 1065-1079.

Gross, O. (1956). *A Class of Discrete-Type Minimization Problems.* RAND Corporation, RM-1655-PR, Santa Monica, CA.

Hadley, G. and T. M. Whitin (1963). *Analysis of Inventory Systems.* Prentice-Hall, Inc., Englewood Cliffs, N.J.

Harris, F. (1915). *Operations and Cost.* (Factory Management Series.) A. W. Shaw Co., Chicago, IL.

Higa, I., A. Feyerherm, and A. Machado (1975). Waiting Time in an (S - 1, S) Inventory System, *Operations Research* 23, 674-680.

Hillier, F. S. and G. J. Lieberman (1980). *Introduction to Operations Research.* Holden-Day, Inc., San Francisco, CA.

Hillestad, R .J. and M. J. Carrillo (1980) *Models and Techniques for Recoverable Item Stockage When Demand and the Repair Process are Nonstationary - Part I: Performance Measurement.* RAND Corporation, N-1482-AF, Santa Monica, CA.

Hodges, J.S. (1985). *Modeling the Demand for Spare Parts: Estimating the Variance-to-Mean Ratio and Other Issues.* RAND Corporation, N-2086-AF, Santa Monica, CA.

Hoel, P.G. (1962). *Introduction to Mathematical Statistics.* Third Edition. John Wiley & Sons, Inc., NY. 255-258

Howell, Lawrence D., Capt. (1978). *A Method for Adjusting Maintenance Forecasts to Account for Planned Sortie Lengths.* ASD-TR-78-26M, Wright-Patterson AFB, OH.

Isaacson, K. E., P. Boren, C. L. Tsai, and R. Pyles (1988). *Dyna-METRIC Version 4 : Modeling Worldwide Logistics Support of Aircraft Components.* RAND Corporation, R-3389-AF, 95-96, Santa Monica, CA.

Kaplan, A. J. (1969). *Economic Retention Limits.* Inventory Research Office, U.S. Army Logistics Management Center, Fort Lee, VA.

Kaplan, A. J. (1980). *Mathematics of SESAME Model Stockage Models.* AMSAA Army Inventory Research Office, Philadelphia, PA.

Kaplan, A. J. (1989). Incorporating Redundancy Considerations into Stockage Models. *Naval Research Logistics* 36, 625-638.

Kline, R .C. and C. C. Sherbrooke (1991). *The M-Spare Model (Multiple Spares Prioritization and Availability to Resource Evaluation) Analysts and Users Guide).* Logistics Management Institute, Washington, D.C. Report NS901R2.

Kruse, K. C. (1979). *An Exact N Echelon Inventory Model: The Simple Simon Method.* U.S. Army Research Office, Technical Report TR 79-2.

Lee, Hau L. (1987). A Multi-Echelon Inventory Model for Repairable Items with Emergency Lateral Transshipments. *Management Science* 31, 1302-1316.

Makridakis, S., and S. Hibon (1979). *Accuracy of Forecasting: An Empirical Investigation, Journal of the Royal Statistical Society, Series A.* 142, Part 2, 119.

Muckstadt, J. (1973). A Model for a Multi-Item, Multi-Echelon, Multi-Indenture Inventory System. *Management Science* 20, 472-481.

Muckstadt, J. (1982). *A Multi-Echelon Model for Indentured, Consumable Items.* Tech. Report 5485, School of Operations Research, Cornell University.

O'Malley, T. J. (1983). *The Aircraft Availability Model: Conceptual Framework and Mathematics.* Logistics Management Institute, Washington, D.C.

Palm, C. (1938). Analysis of the Erlang Traffic Formulae for Busy-Signal Arrangements. *Ericsson Technics.* 4, 39-58.

Parzen, E. (1962). *Stochastic Processes.* Holden-Day, San Francisco, 55.

Presutti, V.J. and R.C. Trepp (1970). More Ado About Economic Order Quantities (EOQ). *Naval Research Logistics.* 17, 243-251.

Robbins, H. (1964). The Empirical Bayes Approach to Statistical Decision Problems. *Annals of Mathematical Statistics.* 35, 1-20.

Rodriguez, C. M. and K. Downer (1997). *VMetric Spare Parts Optimization Model Validation.* U.S. Coast Guard Research and Development Center, CG-D-06-98, Groton, Ct.

Schultz, C. R. (1990). On the Optimality of the (S - 1, S) Policy. *Naval Research Logistics* 37, 715-723.

Schwarz, L. B. (1981). Multi-Level *Production/Inventory Control Systems: Theory and Practice.* North-Holland, Amsterdam

Shaw, C. C., Col. (1981). *Saber Sustainer Briefing.* AF/SAGM, Wright-Patterson AFB, OH.

Sherbrooke, C. C. (1966). Generalization of a Queueing Theorem of Palm to Finite Populations. *Management Science* 12, 907-908.

Sherbrooke, C. C. (1968). METRIC: A Multi-Echelon Technique for Recoverable Item Control. *Operations Research* 16, 122-141.

Sherbrooke, C. C. (1969). *Improved Decision Rules for the Empirical Bayes Problem.* Ph.D. Dissertation, Department of Biostatistics, University of California, Los Angeles.

Sherbrooke, C. C. (1971). An Evaluator for the Number of Operationally Ready Aircraft in a Multi-Level Supply System. *Operations Research* 19, 618-635.

Sherbrooke, C. C. (1975). Waiting Time in an (S - 1, S) Inventory System - Constant Service Time Case, *Operations Research* 23, 819-820.

Sherbrooke, C. C. (1984). *Estimation of the Variance-to-Mean Ratio for AFLC Recoverable Items.* Sherbrooke & Associates, Potomac, MD.

Sherbrooke, C. C. (1986). VARI-METRIC: Improved Approximations for Multi-Indenture, Multi-Echelon Availability Models. *Operations Research* 34, 311-319.

Sherbrooke, C. C. (1987). *Evaluation of Demand Prediction Techniques.* Logistics Management Institute, Washington, D.C. Report AF601R1

Sherbrooke, C. C. (1988). *Backorder Estimation Under Multiple Failures of Lower Indenture Items.* Logistics Management Institute, Washington, D.C. Report AF801R1.

Sherbrooke, C. C. (1997). *Using Sorties vs. Flying Hours to Predict Aircraft Spares Demand.* Logistics Management Institute, Washington, D.C. Report AF501LN1.

Simon, R. M. (1971). Stationary Properties of a Two-Echelon Inventory Model for Low-Demand Items. *Operations Research* 19, 761-773.

Slay, F. M. (1984). *VARI-METRIC: An Approach to Modelling Multi-Echelon Resupply when the Demand Process is Poisson with a Gamma Prior.* Logistics Management Institute, Washington, D.C. Report AF301-3.

Slay, F. M. (1986). *Lateral Resupply in a Multi-Echelon Inventory System.* Logistics Management Institute, Washington, D.C. Report AF501-2.

Slay, F. M. and R. M. King. (1987). *Prototype Aircraft Sustainability Model.* Logistics Management Institute, Washington, D.C. Report AF601-R2.

Slay, F. M. and C. Sherbrooke. (1988). *The Nature of the Aircraft Component Failure Process: A Working Note.* Logistics Management Institute, Washington, D.C. Report IR701R1.

Slay, F. M., C. Sherbrooke, and D. K. Peterson (1996). Predicting Wartime Demand for Aircraft Spares. Air Force Journal of Logistics, Vol XX, No. 2 16-22. Maxwell AFB, Alabama.

Smith, J., W. Fisher, and J. Heller (1972). *Measurements of Military Essentiality.* Logistics Management Institute, Washington, D.C.

Stevens, R.J. and J. M. Hill (1973). *Estimating the Variance-to-Mean Ratio for Recoverable Items in the ALS Marginal Analysis Algorithms.* Working Paper 49. System Studies Branch, Air Force Logistics Command, Wright-Patterson AFB, OH.

Svoronos, A. P. (1986). *A General Framework for Multi-Echelon Inventory and Production Control Problems.* Ph.D. Dissertation, Columbia University, 62.

Svoronos, A. P. and P. Zipkin (1988). Estimating the Performance of Multi-Level Inventory Systems. *Operations Research* 36, 57-72.

Syski, R. (1986). *Introduction to Congestion Theory in Telephone Systems.* Second Edition, North-Holland, Amsterdam, 270-271.

TFD Solutions, Inc. (2000). *C-5 Sparing Analysis Demonstration.* TFD Solutions, Inc., Monterey, CA., 15 August 2000.

Winkler, R. and S. Makridakis (1983). The Combination of Forecasts. *Journal of the Royal Statistical Society, Series A.* 146, Part 2, 150-157.

Index

Early Titles in the

INTERNATIONAL SERIES IN
OPERATIONS RESEARCH & MANAGEMENT SCIENCE
Frederick S. Hillier, Series Editor, *Stanford University*

Saigal/ *A MODERN APPROACH TO LINEAR PROGRAMMING*
Nagurney/ *PROJECTED DYNAMICAL SYSTEMS & VARIATIONAL INEQUALITIES WITH APPLICATIONS*
Padberg & Rijal/ *LOCATION, SCHEDULING, DESIGN AND INTEGER PROGRAMMING*
Vanderbei/ *LINEAR PROGRAMMING*
Jaiswal/ *MILITARY OPERATIONS RESEARCH*
Gal & Greenberg/ *ADVANCES IN SENSITIVITY ANALYSIS & PARAMETRIC PROGRAMMING*
Prabhu/ *FOUNDATIONS OF QUEUEING THEORY*
Fang, Rajasekera & Tsao/ *ENTROPY OPTIMIZATION & MATHEMATICAL PROGRAMMING*
Yu/ *OR IN THE AIRLINE INDUSTRY*
Ho & Tang/ *PRODUCT VARIETY MANAGEMENT*
El-Taha & Stidham/ *SAMPLE-PATH ANALYSIS OF QUEUEING SYSTEMS*
Miettinen/ *NONLINEAR MULTIOBJECTIVE OPTIMIZATION*
Chao & Huntington/ *DESIGNING COMPETITIVE ELECTRICITY MARKETS*
Weglarz/ *PROJECT SCHEDULING: RECENT TRENDS & RESULTS*
Sahin & Polatoglu/ *QUALITY, WARRANTY AND PREVENTIVE MAINTENANCE*
Tavares/ *ADVANCES MODELS FOR PROJECT MANAGEMENT*
Tayur, Ganeshan & Magazine/ *QUANTITATIVE MODELS FOR SUPPLY CHAIN MANAGEMENT*
Weyant, J./ *ENERGY AND ENVIRONMENTAL POLICY MODELING*
Shanthikumar, J.G. & Sumita, U./ *APPLIED PROBABILITY AND STOCHASTIC PROCESSES*
Liu, B. & Esogbue, A.O./ *DECISION CRITERIA AND OPTIMAL INVENTORY PROCESSES*
Gal, T., Stewart, T.J., Hanne, T. / *MULTICRITERIA DECISION MAKING: Advances in MCDM Models, Algorithms, Theory, and Applications*
Fox, B.L. / *STRATEGIES FOR QUASI-MONTE CARLO*
Hall, R.W. / *HANDBOOK OF TRANSPORTATION SCIENCE*
Grassman, W.K. / *COMPUTATIONAL PROBABILITY*
Pomerol, J-C. & Barba-Romero, S. / *MULTICRITERION DECISION IN MANAGEMENT*
Axsäter, S. */ INVENTORY CONTROL*
Wolkowicz, H., Saigal, R., & Vandenberghe, L. / *HANDBOOK OF SEMI-DEFINITE PROGRAMMING: Theory, Algorithms, and Applications*
Hobbs, B.F. & Meier, P. / *ENERGY DECISIONS AND THE ENVIRONMENT: A Guide to the Use of Multicriteria Methods*
Dar-El, E. / *HUMAN LEARNING: From Learning Curves to Learning Organizations*
Armstrong, J.S. / *PRINCIPLES OF FORECASTING: A Handbook for Researchers and Practitioners*
Balsamo, S., Personé, V., & Onvural, R./ *ANALYSIS OF QUEUEING NETWORKS WITH BLOCKING*
Bouyssou, D. et al. / *EVALUATION AND DECISION MODELS: A Critical Perspective*
Hanne, T. / *INTELLIGENT STRATEGIES FOR META MULTIPLE CRITERIA DECISION MAKING*
Saaty, T. & Vargas, L. / *MODELS, METHODS, CONCEPTS and APPLICATIONS OF THE ANALYTIC HIERARCHY PROCESS*
Chatterjee, K. & Samuelson, W. / *GAME THEORY AND BUSINESS APPLICATIONS*
Hobbs, B. et al. / *THE NEXT GENERATION OF ELECTRIC POWER UNIT COMMITMENT MODELS*
Vanderbei, R.J. / *LINEAR PROGRAMMING: Foundations and Extensions, 2nd Ed.*
Kimms, A. / *MATHEMATICAL PROGRAMMING AND FINANCIAL OBJECTIVES FOR SCHEDULING PROJECTS*
Baptiste, P., Le Pape, C. & Nuijten, W. / *CONSTRAINT-BASED SCHEDULING*
Feinberg, E. & Shwartz, A. / *HANDBOOK OF MARKOV DECISION PROCESSES*: *Methods and Applications*

**** A list of the more recent publications in the series is at the front of the book ****