INTERNATIONAL SERIES IN OPERATIONS
RESEARCH AND MANAGEMENT SCIENCE

# Building Intuition

## Insights from Basic Operations Management Models and Principles

Dilip Chhajed
Timothy J. Lowe
*Editors*

Springer

# Building Intuition

## Insights From Basic Operations Management Models and Principles

*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*

Dilip Chhajed • Timothy J. Lowe
Editors

# Building Intuition

## Insights From Basic Operations Management Models and Principles

Springer

*Editors*
Dilip Chhajed
University of Illinois
Champaign, IL 61822 USA
chhajed@illinois.edu

Timothy J. Lowe
University of Iowa
Iowa City, IA 52242 USA
Timothy-Lowe@uiowa.edu

*Series Editor*
Fred Hillier
Stanford University
Stanford, CA, USA

ISBN 978-0-387-73698-3 e-ISBN 978-0-387-73699-0

Library of Congress Control Number: 2007941122

Printed on acid-free paper

9 8 7 6 5 4 3 2 1

springer.com

# Dedication

*To my wife Marsha, and my children*
*Marc and Carrie*
*-TL*

*To my parents, Aradhana, Avanti and Tej*
*-DC*

# Author Bios

Dilip Chhajed is Professor of Operations Management and the Director of the Master of Science in Technology Management program at the University of Illinois at Urbana-Champaign. Professor Chhajed received his Ph.D. from Purdue University in Management, and also has degrees from University of Texas (Management Information Systems) and Indian Institute of Technology (Chemical Engineering).
He has held visiting professor appointments at Purdue University, International Postgraduate Management Center at the University of Warsaw in Poland, Sogang University in South Korea, and GISMA in Hanover, Germany. He has also served as the Faculty-in-Residence at Caterpillar Logistics Services. His research interests focus on decision problems in operations management and the operations/marketing interface including supply chain management and product/process design. He has authored or co-authored over 35 articles and invited chapters. Professor Chhajed's teaching interests are in Process/Operations Management, Project Management, Process Improvement, and Supply Chain Management. He has taught in the MBA, Executive MBA, and MS in Technology Management programs. He has consulted and participated in executive education in the area of process design and improvement.

Dilip Chhajed

Timothy J. Lowe is the Chester Phillips Professor of Operations Management at the Tippie College of Business, University of Iowa. He formerly served as Director of the College's Manufacturing Productivity Center. He has teaching and research interests in the areas of operations management and management science. He received his BS and MS degrees in engineering from Iowa State University and his Ph.D. in operations research from Northwestern University. He has served as Associate Editor for the journals: *Operations Research, Location Science, TOP*, and *Managerial and_Decision Economics*; as a Senior Editor for *Manufacturing and Service Operations_Management;* and as Departmental Editor (Facilities Design/Location Analysis) for *Transactions of the Institute of Industrial Engineers*. He has held several grants from the National Science Foundation to support his research on location theory. He has published more than 80 papers in leading journals in his field. Professor Lowe has worked as a project/process engineer for the Exxon Corporation, and has served on the faculties of the University of Florida, Purdue University and Pennsylvania State University. At Purdue, he served as the Director of Doctoral Programs and Research for the Krannert Graduate School of Management. He has considerable experience in executive education both in the

U.S., as well as overseas. In addition, he has served as a consultant in the areas of production and distribution for several companies.

Timothy J. Lowe

# Foreword

The year is 2027, the price of quantum computers is falling rapidly, and a universal solver for the leading type of quantum computer is now in its second release. Given any model instance and any well-formulated problem posed on that instance, Universal Solver 2.0 will quickly produce a solution. There are some limitations, of course: the model instance has to be specified using a given repertoire of common mathematical functions within an algebraic framework or, possibly, a system of differential or integral equations, with the data in a particular kind of database, and the problem posed has to be of a standard type: database query, equation solution, equilibrium calculation, optimization, simulation, and a few others.

As a tool for practical applications of operations management and operations research, is this all we need?

I think not. Useful as such a tool would be, we still need a solver or solution process that can explain *why* a solution is what it is, especially when the validity of the solution is not easily verifiable. The big weakness of computations and solvers is that they tell you *what* but not *why*.

Practitioners need to know why a solver gives the results it does in order to arrive at their recommendations. A single model instance—that is, a particular model structure with particular associated data—hardly ever suffices to capture sufficiently what is being modeled. In practical work, one nearly always must solve multiple model instances in which the data and sometimes even the model structure are varied in systematic ways. Only then can the practitioner deal with the uncertainties, sensitivities, multiple criteria, model limitations, etc., that are endemic to real-life applications. In this way, the practitioner gradually figures out the most appropriate course of action, system design, advice, or whatever other work product is desired.

Moreover, if a practitioner cannot clearly and convincingly explain the solutions that are the basis for recommendations—especially to the people who are paying for the work or who will evaluate and implement the recommendations—then it is unlikely that the recommendations will ever come to fruit or that the sponsor will be fully satisfied.

There are two major approaches to figuring out why a model leads to the solutions that it does. One is mainly computational. In the course of solving multiple model instances, as just mentioned, the analyst comes to understand some of the solution characteristics well enough to justify calling them insights into why the solutions are what they are (typically at an aggregate rather than detailed level). These insights can inform much of the thinking that the model was designed to facilitate and can facilitate communicating with others.

The second approach is not primarily computational, but rather is based on developing insights into model behavior by analytical means. Direct analytical study may be possible for very simple (idealized) model structures, but this tends not to be feasible for the kinds of complex models needed for most real applications. Practical studies may have to rely on a deep understanding of greatly simplified models related to the one at hand, or on long experience with similar models. This is an art leading mainly to conjectures about solution characteristics of interest for the fully detailed model, and was the approach taken in the paper of mine that the editors cite in their preface. These conjectures are then subjected to computational or empirical scrutiny, and the ones that hold up can be called insights into why the full model's solutions are what they are.

The importance of this book, in my view, rests partly on its success in teasing out the deep understanding that is possible for some relatively simple yet common model structures, which in turn can be useful for the second approach just sketched, and partly on the sheer expository strength of the individual chapters. The profession can never have too many excellent expositions of essential topics at the foundation of operations management. These are valuable for all the usual reasons—utility to instructors, utility and motivation for students and practitioners, utility to lay readers (perhaps even the occasional manager or engineering leader) curious about developments in fields outside their own expertise, and even utility to researchers who like to accumulate insights outside their usual domain.

Having stressed the *utility* of expositions that communicate the insights attainable by avoiding too many complexities, let me balance that by pointing out how exquisitely *beautiful* the insights of such expositions can be, and also how exquisitely *difficult* such writing is.

Most readers will find a good deal of beauty as well as utility in this book's chapters, and I commend the editors and authors for their efforts.

Arthur Geoffrion
UCLA Anderson School of Management

# Preface

The idea for this book began with a discussion at a professional meeting regarding teaching materials. As educators in schools of business, we each were looking for materials and teaching approaches to motivate students of operations management regarding the usefulness of the models and methods presented in the basic OM course. Our experience has been that many of the basic OM concepts have been "fleshed out" and so deeply developed to the point where basic insights are often lost in the details. Over the years, we both have been heavily influenced by Art Geoffrion's classic article "The Purpose of Mathematical Programming is Insight, Not Numbers," *Interfaces*, 1976. We believe that this principle is fundamental in educating users of the "products" we deliver in the classroom, and so our project—this book—was initiated with a great deal of enthusiasm. Our first task was to enlist the assistance of well-known individuals in the field who have the professional credentials to gain the attention of potential readers, yet are able to tell their story in language appropriate for our target audience. We think you will agree that we have been successful in our choice of authors.

The purpose of this book is to provide a means for making selected basic operations management models and principles more accessible to students and practicing managers. The book consists of several chapters, each of which is written by a well-known expert in the field. Our hope is that this user-friendly book will help the reader to develop insights with respect to a number of models that are important in the study and practice of operations management. We believe that one of the primary purposes of any model is to build intuition and generate insights. Often, a model is developed to be able to better understand phenomena that are otherwise difficult to comprehend. Models can also help in verifying the correctness of an intuition or judgment. As an example, managers may use the SPT (shortest processing time) method to schedule completion of paperwork with the objective of "clearing their desk"— removing as many jobs from their desk as quickly as possible. As it turns out, it can be easily shown that SPT sequencing minimizes average job flow time (see Chap. 1). Thus, in this case, it is comforting to know that the manager's intuition is correct. However, it is also essential to know when (and why!) intuition fails, and a well-structured model should convey this information. In spite of the fact that many educators recognize the intuition-building power of simple models, we are not aware of any existing book that has a focus similar to ours.

As mentioned above, Chap. 1 deals with the shortest process time principle. Chapter 2 contains insight on the knapsack problem—a problem that often arises as a subproblem in more complex situations. The notion of process flexibility, and how to efficiently attain it, is the subject of Chap. 3, while queuing concepts are the subject of Chap. 4. A key relationship between flow rate, flow time, and units in the system—Little's Law—is discussed in Chap. 5. In Chap. 6, the use of the median, as opposed to mean, is shown to be a best choice in certain situations. The newsvendor model, a means of balancing "too much" versus "not enough," is the subject of Chap. 7, while the economic order quantity inventory model is covered in Chap. 8. The pooling principle, a means of mitigating variance, is the topic of the final chapter.

To ensure that the book is accessible by our target audience, the chapters are written with students and managers in mind. Reading the book should help in developing a deeper appreciation for models and their applications. One measure of accessibility is that individuals only vaguely familiar with OM principles should be able to read and comprehend major portions of the book. We sincerely hope that the book will meet this test.

This book should appeal to three major audiences: (a) teachers of introductory courses in OM, (b) students who are taking one of their first courses in OM, and (c) managers who face OM decisions on a regular basis.

As professors who have considerable experience in teaching OM, we have found that students value insights gained by the models and tools that are the subject of this book. In addition, early in our careers we experienced a certain level of "discomfort" in teaching some of these models. This discomfort arose because as teaching "rookies" we lacked the maturity and experience to do proper justice to the material. Thus, we hope that the background and examples provided by the book will be of considerable help to "new" teachers.

Finally, we hope that the book will also appeal to those managers who believe that decision technology tools can be brought to bear on the problems they face.

Although each chapter of this book treats a different fundamental OM concept, we have made every effort to have a uniform writing style and (as much as possible) consistency in notation, etc. With this in mind, the book can be used in its entirety in an OM course. Alternatively, individual chapters can be used in a stand-alone situation since material does not "build" progressively through the book. For our managerial audience, we see the book as an excellent reference source.

We sincerely hope that you will find the book useful and that it will be a valuable addition to your personal library.

Tim Lowe and Dilip Chhajed

# Contents

# Chapter 1
# Sequencing: The Shortest Processing Time Rule

**Kenneth R. Baker**
**Dartmouth College**

*The problem of sequencing arises frequently in many areas of business and engineering. Shortest-first sequencing (or one of its variants) has proven to be optimal in a number of problem areas.*

## Introduction

Thanks to a day's exposure at the annual professional conference, Alison has received four requests for design work. Based on her experience, she can estimate how long it would take her to deliver each design, noting that some of the jobs were short ones and others were longer. In addition, Alison's working style is to focus on one job at a time, postponing any thought about the others until the current one is finished. Nevertheless, as she thinks about the situation, she realizes that her four customers will get different impressions of her responsiveness depending on the order in which she completes the designs. In other words, it seems to make a difference what sequence she chooses for the jobs.

### *What is Alison's Sequencing Problem?*

Let's start by acknowledging that Alison is indeed faced with a decision. She could work on her four jobs in any one of several possible sequences. Depending on what order she chooses, at least one of the customers will get a response almost immediately, while the rest of the customers will experience various delays while they wait for other work to be completed. When different customers experience different responses, it helps if we can find a suitable measure to evaluate how effectively the entire set of customers is served.

For example, suppose Alison's predictions of the job times are as follows. The times indicate how many days must be dedicated to each job.

D. Chhajed and T.J. Lowe (eds.) *Building Intuition*: *Insights From Basic Operations Management Models and Principles*.
doi:10.1007/978-0-387-73699-0, © Springer Science + Business Media, LLC 2008

| | | | | |
|---|---|---|---|---|
| Job number | 1 | 2 | 3 | 4 |
| Work time | 4 | 2 | 9 | 7 |

From the time Alison starts, the entire set of jobs will take 4 + 2 + 9 + 7 = 22 days to complete. Let's examine how individual customers might experience delays.

Suppose that the jobs are sequenced by job number (1-2-3-4). Then we can build the following table to trace the implications.

| | | | | |
|---|---|---|---|---|
| Sequence position | 1 | 2 | 3 | 4 |
| Job number | 1 | 2 | 3 | 4 |
| Work time | 4 | 2 | 9 | 7 |
| Completion time | 4 | 6 | 15 | 22 |

The completion time of any customer is the completion time of the previous job in sequence, plus the time for the waiting customer's job. A graphical representation of the sequence is shown in Exhibit 1.1.

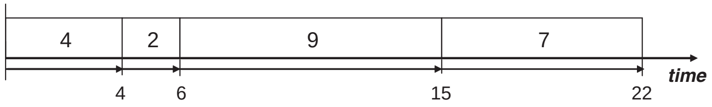


*Exhibit 1.1*

The four completion times measure the four delay times experienced by the customers. One simple way to measure the effectiveness of the entire schedule is simply to add these numbers: 4 + 6 + 15 + 22 = 47. Some other sequence (for example, 4-3-2-1) would achieve a different measure of effectiveness (63), as calculated from the table below.

| | | | | | |
|---|---|---|---|---|---|
| Sequence position | 1 | 2 | 3 | 4 | |
| Job number | 4 | 3 | 2 | 1 | |
| Work time | 7 | 9 | 2 | 4 | |
| Completion time | 7 | 16 | 18 | 22 | 7 + 16 + 18 + 22 = 63 |

In other words, the total of the delay times experienced by individual customers is 47 in the former case and 63 in the latter. Thus, different sequences can lead to different measures of effectiveness: in sum, it *does* make a difference which sequence Alison chooses. Moreover, because smaller response times are more desirable than larger ones, Alison would prefer a sequence with a measure of 47 to one with a measure of 63. In fact, Alison would like to find the sequence that *minimizes* this measure of effectiveness.

## *What is the Sequencing Problem?*

We can begin to generalize from the example of Alison's decision. A sequencing problem contains a collection of jobs. In our example, there are 4 jobs; in general, we could have *n* jobs. Since the timing of those jobs is usually critical to the problem, a sequencing problem also specifies the time required to carry out each of the given jobs. In our example, there are job times of 4, 2, 9, and 7. In general, we could associate the processing time $t_i$ with the *i*th job. Thus, jobs and job times constitute the minimal description of a sequencing problem. As long as the job times are not all identical, different sequences will exhibit different scheduling-related properties. These differences lead us to specify a measure of effectiveness, or *objective*, which allows us to compare sequences and ultimately find the best one. In our example, the objective is the sum of four completion times. If we had *n* jobs, we could write the objective as $C_1 + C_2 + \dots + C_n$, where $C_i$ represents the completion time of job *i* (or, equivalently, the delay time experienced by the corresponding customer). The usual practice is to adopt the shorthand $\Sigma C$ to represent the sum of the jobs' completion times.

It is possible to generalize further and consider objectives that are more complicated functions of the completion times, but we shall look at that situation later. For now, an objective that consists of the sum of the completion times adequately captures one of the most common scheduling concerns—the response times experienced by a set of customers.

Thus, the sequencing problem starts with given information (*n* and the $t_i$-values) and identifies an objective. The decision problem is then to choose the best sequence. In our example, we can find the best sequence by listing every possible sequence and for each one, calculating the sum of completion times. The list of possible sequences (along with the objective value for each one) is shown in Exhibit 1.2.

| Sequence | | | | Objective |
|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 47 |
| 1 | 2 | 4 | 3 | 45 |
| 1 | 3 | 2 | 4 | 54 |
| 1 | 3 | 4 | 2 | 59 |
| 1 | 4 | 2 | 3 | 50 |
| 1 | 4 | 3 | 2 | 57 |
| 2 | 1 | 3 | 4 | 45 |
| 2 | 1 | 4 | 3 | 43 |
| 2 | 3 | 1 | 4 | 50 |
| 2 | 3 | 4 | 1 | 53 |
| 2 | 4 | 1 | 3 | 46 |
| 2 | 4 | 3 | 1 | 51 |
| 3 | 1 | 2 | 4 | 59 |
| 3 | 1 | 4 | 2 | 64 |
| 3 | 2 | 1 | 4 | 57 |
| 3 | 2 | 4 | 1 | 60 |
| 3 | 4 | 1 | 2 | 67 |
| 3 | 4 | 2 | 1 | 65 |
| 4 | 1 | 2 | 3 | 53 |
| 4 | 1 | 3 | 2 | 60 |
| 4 | 2 | 1 | 3 | 51 |
| 4 | 2 | 3 | 1 | 56 |
| 4 | 3 | 1 | 2 | 65 |
| 4 | 3 | 2 | 1 | 63 |

*Exhibit 1.2*

By comparing all twenty-four sequences, we can verify that the best sequence in our example is 2-1-4-3, with an objective of 43.

In general, we could imagine writing down all of the possible sequences and then computing the objective for each one, eventually identifying the best value. However, before we set out on such a task, we should take a moment to anticipate how much work we are taking on. The number of possible sequences is easy to count: there are *n* possibilities for choosing the first job in sequence, for each choice there are $(n - 1)$ possibilities for the second job in sequence, then $(n - 2)$ possibilities for the third job, and so on. The total number of sequences is therefore $(n)\times(n - 1)\times(n - 2) \times\ldots \times (2)\times(1) = n!$ In our example, this came to 4! = 24 possible sequences, as listed in Exhibit 1.2. In general, if *n* were, say, 20, or perhaps 30, listing all the possibilities might become too tedious to do by hand and even too time-consuming to do with the aid of a computer. We should like to have a better way of finding the best sequence, because listing all the possibilities will not always be practical. Besides, listing all the possibilities amounts to no more than a "brute force" method of solving the problem. We would much prefer to develop some general insight about the nature of the problem and then apply that insight in order to produce a solution

## Pairwise Interchanges and Shortest-First Priority

We can often learn a lot about solving a sequencing problem from a simple interchange mechanism: take two adjacent jobs somewhere in a given sequence and swap them; then determine whether the swap led to an improvement. If it did, keep the new sequence and try another swap; if not, just go back to the original sequence and try a different swap.

**Building Intuition** Sequencing jobs using Shortest Processing Time rule (shortest job first) minimizes not only the sum of completion times but also minimizes average completion time, maximizes the number of jobs that can be completed by a pre-specified deadline, minimizes the sum (and average) of wait times, and minimizes the average inventory in the system. Although the delay criterion and the inventory criterion may appear to be quite different, they are actually different sides of the same coin: both are optimized by SPT sequencing. Thus, whether a manager focuses on customer delays or on unfinished work in progress, or tries to focus on both at once, shortest-first sequencing is a good idea.

When jobs have different priorities or inventory costs, the job time must be divided by the weight of the job where the weight may represent priority or cost. These weighted times can then be used to arrange the jobs.

For all these problems suppose we start with any sequence and swap two adjacent jobs if it leads to improvement. Performing all adjacent pairwise interchanges (APIs) will always lead us to the optimal sequence.

Let us illustrate this mechanism with our example. Suppose we start with the sequence 1-3-2-4, with an objective of 54. Try swapping the first two jobs in sequence, yielding the sequence 3-1-2-4, which has an objective of 59. In other words, the swap made things worse, so we go back to the previous sequence and swap the second and third jobs in sequence. This swap yields the sequence 1-2-3-4, with an objective of 47. This time, the swap improved the objective, so we'll keep the sequence. Returning to the beginning of the sequence (because the first pair of jobs has been altered), we again try swapping the first two jobs, yielding the sequence 2-1-3-4, with an objective of 45. Now, we don't have to revisit the first two jobs in sequence because we just confirmed that they appear in a desirable order. Next, we go to the second and third jobs; they, too, appear in a desirable order. Next, we go to the third and fourth jobs, where the swap yields 2-1-4-3, with an objective of 43. From Exhibit 1.2 we recognize this sequence as the optimal one.

Testing the interchange ("swap") of two adjacent jobs, to see whether an improvement results, gives us a mechanism to start with any sequence and to look for improvements. Searching for improvements this way cannot leave us worse off, because we can always revert to the previous sequence if the swap does not lead to an improvement. In our example, a series of adjacent pairwise interchanges led us to an optimal sequence; moreover, we evaluated fewer sequences than were necessary with the "complete enumeration" approach of Exhibit 1.2 In general, we might wonder whether adjacent pairwise interchanges (APIs) will always lead us to the optimal sequence.

## APIs in General

Imagine that we have a sequence on hand that contains *n* jobs. Somewhere, possibly in the middle of that sequence, is a pair of jobs, *j* and *k*, with *k* following immediately after *j*. Suppose we swap jobs *j* and *k* and trace the consequences in general terms. See Exhibit 1.3 for a display of the swap.

Let S denote the original sequence on hand, and let S' refer to the sequence after the swap is made. When we write $C_i(S)$, we refer to the completion time of job *i* in

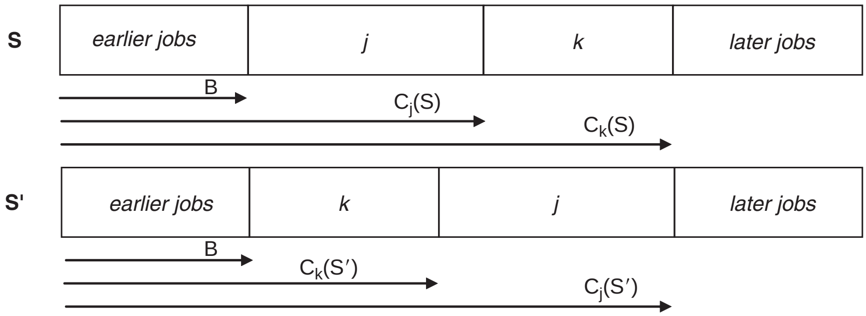


*Exhibit 1.3*

schedule S, and when we write Σ *C*(S), we refer to the sum of completion times in schedule S. Note that swapping adjacent jobs *j* and *k* does not affect the completion of any other jobs, so we can write *B* to represent the completion time of the job immediately before *j* in schedule S, and we can write *C* to represent the sum of the completion times of all jobs other than *j* and *k*. Note that *B* and *C* are unchanged by the swap. Referring to Exhibit 1.3, we find that

$$\Sigma C(S) = \mathrm{C}+\mathrm{C}_j(S) + \mathrm{C}_k(S) = \mathrm{C}+(\mathrm{B}+\mathrm{t}_j) + (\mathrm{B}+\mathrm{t}_j+\mathrm{t}_k), \text{ and}$$

$$\Sigma C(S') = \mathrm{C}+\mathrm{C}_j(\mathrm{S}') + \mathrm{C}_k(\mathrm{S}') = \mathrm{C}+(\mathrm{B}+\mathrm{t}_k+\mathrm{t}_j) + (\mathrm{B}+\mathrm{t}_k).$$

Then the difference between the two objectives becomes:

$$\Delta=\Sigma \mathrm{C}(S') - \Sigma C(S') = \mathrm{t}_j - \mathrm{t}_k.$$

Since Δ > 0 if the swap improves the objective, it follows that there will be an improvement whenever $t_k < t_j$. In other words, the swap improves the objective if it places the shorter job first.

The implication of this result is far reaching. If we were to encounter *any* sequence in which we could find an adjacent pair of jobs with the longer job first, we could swap the two jobs and thereby create a sequence with a smaller value of Σ*C*. Thus, the only sequence in which improvement is not possible would be a sequence in which the first job in any adjacent pair is shorter (or at least no longer) than the following job. In other words, the optimal sequence is one that processes the jobs in the order of Shortest Processing Time, or SPT.

***Property 1-1.*** *When the objective is the sum of completion times, the minimum value is achieved by sequencing the jobs in the order of Shortest Processing Time.*

This property articulates the virtue of shortest-first sequencing. For any set of *n* jobs, shortest-first priority leads to a sequence that minimizes the value of Σ*C*. Thus, whenever the quality of a sequence is measured by the sum of completion times, we can be sure that the appropriate sequencing rule is *shortest first*.

Armed with Property 1-1, we can easily find the optimal sequence for our example. Arranging the jobs in shortest-first order, we can immediately construct the sequence 2-1-4-3 and then calculate the corresponding sum of completion times as 43. There is no need to enumerate all 24 sequences, as we did in Exhibit 1.2, nor is there even a need to choose an arbitrary schedule and begin applying pairwise interchanges in search of improvements, as illustrated in Exhibit 1.3. The power of Property 1-1 is that we can avoid searching and improvement efforts and construct the optimal sequence directly.

## Other Properties of SPT

Recall that a sequencing problem is defined by a set of jobs and their processing times, along with an objective. The choice of objective is a key step in specifying the problem, and the use of the sum of completion times may sound specialized. However, SPT is optimal for other measures as well.

First, note that Property 1-1 applies to the objective of the average completion time. The average would be computed by dividing the sum of completion times by the (given) number of jobs. In our example, that would mean dividing by 4, yielding an optimal average of 10.75; in general, that would mean dividing by the constant *n*. Therefore, whether we believe that the *sum* of the completion times or the *average* of the completion times more accurately captures the measurement of schedule effectiveness, there is no difference when it comes to the sequencing decision: SPT is optimal.

Second, consider the objective of completing as many jobs as possible by a pre-specified deadline. No matter what deadline is specified, SPT sequencing maximizes the number of jobs completed by that time. This might not be surprising to anyone who has tried to squeeze as many different tasks as possible into a finite amount of time. Choosing the small tasks obviously works better than choosing the large tasks, if our objective is to complete as many as possible. Taking this observation to its limit, it follows that a shortest-first sequence maximizes the number of tasks completed.

Next, consider the completion time as a measure of the delay experienced by an individual customer. We could argue that the total delay has two parts: the wait for work to start and the time of the actual work itself. The processing time is a function of the order requested by the customer and is therefore likely to be under the customer's control. The wait until work starts depends on the sequence chosen. Thus, we might prefer to take as an objective the sum of the wait times. Although this seems at first glance to be a different problem, it is not hard to see that SPT is optimal for the sum of the wait times (and the average of the wait times) as well.

Finally, consider a scheduling objective oriented not to the delays experienced by customers but rather to the number of jobs waiting in the system at any time. Let *N*(*t*) represent the number of jobs in process (not yet completed) at time *t*. If we

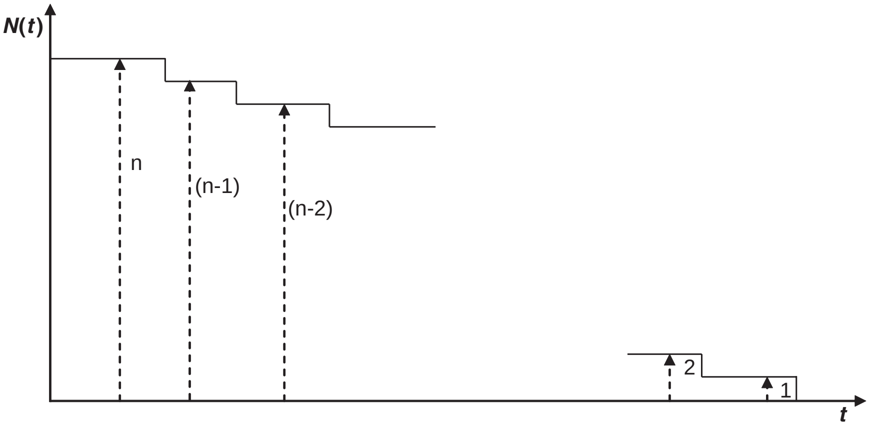


*Exhibit 1.4*

were to draw a graph of the function $N(t)$, it would resemble the stair-step form of Exhibit 1.4.

Now suppose that the jobs are sequenced in numerical order (1, 2, …, $n$) as in Exhibit 1.4. The function $N(t)$ is level at a height of $n$ for a time interval equal to the processing time of the first job, or $t_1$. Then the function drops to $(n – 1)$ and remains level there for a time interval of $t_2$. Then the function drops to $(n – 2)$ and continues in this fashion, until it eventually drops to zero at time $\Sigma\ t_k$. In fact, any sequence leads to a graph that starts out level at a height of $n$ and eventually drops to zero at time $\Sigma\ t_k$. However, a desirable sequence would be one that tends to keep $N(t)$ relatively low. To create a specific objective for this goal, we form the *time average* of the function $N(t)$, which we can think of as the average inventory. That is, we weight each value of $N(t)$ by the time this value persists, the values of $N(t)$ being $n$, $(n – 1)$, $(n – 2)$, …, 2, 1. In this fashion, we form a weighted sum composed of the following pairwise products:

$$\mathrm{n}\times[\text{length of time } N(\mathrm{t}) = \mathrm{n}]$$

$$(\mathrm{n}–1)\times[\text{length of time } N(\mathrm{t}) = (\mathrm{n}–1]$$

$$(\mathrm{n}–2)\times[\text{length of time } N(\mathrm{t}) = (\mathrm{n}–2)]$$

$$\ldots$$

$$2\times[\text{length of time } N(\mathrm{t}) = 2]$$

$$1\times[\text{length of time } N(\mathrm{t})=1].$$

The first of these products is the area of the vertical strip in Exhibit 1.4 corresponding to the length of time that there are $n$ jobs in the system. The next product is the area corresponding to the length of time that there are $(n – 1)$ jobs in the system, and so on. When we sum all the pairwise products (i.e., the areas of the vertical strips), we obtain the area under the graph of the $N(t)$ function. Then we divide this sum-of-products by the total time required by the schedule, or $\Sigma t_i$, in order to produce the average inventory.

It is not difficult to show that the area under the graph, and therefore the average inventory, is minimized by SPT. Perhaps the quickest way to see this result is to use the API illustrated in Exhibit 1-3. Suppose that the completion of the "earlier jobs" in the exhibit leaves $J$ jobs in inventory. In schedule S, that leaves $J$ jobs during the processing of job $j$ and $(J – 1)$ jobs during the processing of job $k$. In schedule S', these numbers are reversed. Now let $A$ represent the contribution of the earlier and later jobs to the area under the graph. (This contribution is unaffected by the interchange of jobs $j$ and $k$.) Then we can write:

$$Area(S) = A + J\mathrm{t}_j + (J–1)\mathrm{t}_k, \text{ and}$$

$$Area(S') = A + J\mathrm{t}_k + (J–1)\mathrm{t}_j.$$

The difference between the two objectives becomes:

$$\Delta = Area(S) – Area(S') = t_j – t_k.$$

The area is smaller whenever $\Delta > 0$, so there will be an improvement whenever $t_k < t_j$. It follows that the minimum possible area (corresponding to the minimum possible average inventory) is achieved by the SPT sequence.

This derivation shows that shortest-first sequencing minimizes average inventory as well as average completion time under some very specialized conditions. For example, the job processing times were assumed to be known with certainty, and all jobs were available simultaneously. However, the intimate relationship between average inventory and average completion time holds under many other circumstances. See the coverage of Little's Law in Chap. 5 for a broader picture of this relationship.

The dual benefits of SPT sequencing are fortuitous if we think in terms of the criteria that often guide managers of manufacturing or service systems as they strive to satisfy a given set of customer orders. One type of concern is the response time that customers experience. Customers would like delays in satisfying their orders to be short. As we have seen, if we quantify this concern with the average completion time objective, an optimal sequencing strategy is shortest-first. Another type of concern is the pile of unfinished orders in inventory, which managers would like to keep small. If we quantify this concern with the average inventory objective, then shortest-first remains an optimal strategy. In the context of our simple sequencing problem, at least, the delay criterion and the inventory criterion may sound quite different, but they are actually different sides of the same coin: both are optimized by SPT sequencing. Thus, whether a manager focuses on customer delays or on unfinished work in progress, or tries to focus on both at once, shortest-first sequencing is a good idea.

## A Generalization of SPT

The shortest-first insights listed in the previous section are useful principles for making decisions about simple sequencing problems. In practice, there is sometimes a complicating factor: long jobs may often be important ones. Strict priority in favor of the shortest job would often conflict with importance. How do we reconcile this conflict? In other words, how do we harness the virtues of shortest-first sequencing while acknowledging that jobs reflect different levels of importance?

To answer this question, we'll have to go back to our basic problem statement and enrich it appropriately. To incorporate notions of importance, each job must have an importance weighting as well as a processing time. Thus, let's take $w_i$ to be a positive number representing the relative importance (or *weight*) of the *i*th job. (We can think of the values of $w_i$ as lying on a scale between zero and one, but the scale usually doesn't matter.) Now the statement of our sequencing problem contains *n* jobs, each with a given processing time and weight.

Next, we need to specify an objective that is consistent with the use of weights. An obvious candidate is the weighted sum of completion times. In symbols, the weighted sum is $w_1C_1 + w_2C_2 + \ldots + w_nC_n$. This means that, as a result of sequencing choices, each job *i* experiences a completion time equal to $C_i$, and we form the objective by multiplying each completion time by its corresponding weight. Our objective is to minimize the sum of these values. Shorthand for the weighted sum of completion times is $\Sigma wC$.

To develop a sequencing rule suitable for the weighted objective, we can again investigate adjacent pairwise interchanges. As in Exhibit 1.3, we consider a pair of adjacent jobs, *j* and *k*, somewhere in sequence S, and we swap those two jobs to form the sequence S′. This time, we let *C* denote the contributions of the other jobs (excluding *j* and *k*) to the objective, a contribution that is unaffected by the swap, and as before, we let *B* denote the time required to complete all the jobs preceding *j* and *k*. Then

$$\Sigma wC(S)=C+w_jC_j(S)+w_kC_k(S)=C+w_j(B+t_j)+w_k(B+t_j+t_k), \text{ and}$$

$$\Sigma wC(S')=C+w_jC_j(S')+w_kC_k(S')=C+w_j(B+t_k+t_j)+w_k(B+t_k).$$

Then the difference between the two objectives becomes:

$$\Delta = \Sigma C_i(S)–\Sigma C_i(S') = w_kt_j–w_jt_k.$$

Since $\Delta > 0$ if the swap improves the objective, it follows that there will be an improvement if $w_jt_k < w_kt_j$. We can also write this condition as $t_k/w_k < t_j/w_j$. In other words, the swap improves the objective if it places first the job with the smaller *ratio* of processing time to weight. In other words, we can weight each job's processing time by dividing it by the job's weight, then we can sequence the jobs according to the shortest *weighted* processing time (SWPT).

***Property 1-2.*** *When the objective is the sum of weighted completion times, the minimum value is achieved by sequencing the jobs in the order of Shortest Weighted Processing Time.*

Property 1-2 reveals that to reconcile the conflict between processing time and importance weight, we should use the ratio of time to weight. Viewed another way, the implication is that we should first scale the processing times (divide each one by its corresponding weight) and then apply the shortest-first principle to the scaled processing times.

Property 1-2 also applies to situations where the concern is focused on inventory, but where jobs represent different costs. A common criterion in practice is the minimization of average inventory *cost* rather than average number of inventory items. Suppose that each job *i* has an associated cost, $w_i$. Then a suitable objective could be the average inventory value, where items are valued at cost. A variation on this objective

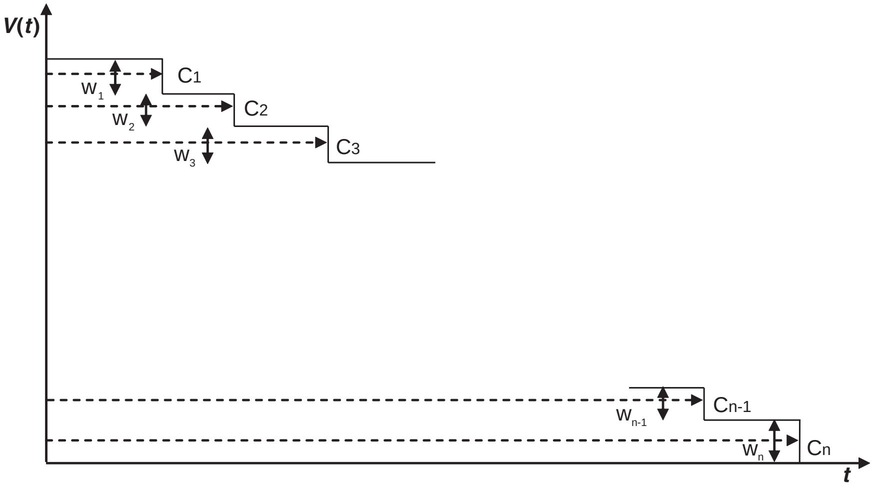


*Exhibit 1.5*

is to value items at some percentage of cost that reflects the opportunity cost of carrying stock. At the outset, all jobs are in inventory, so the inventory value is $\Sigma w_i$. As jobs in the sequence are finished, the value of inventory drops, eventually falling to zero.

A graphical representation is shown in Exhibit 1.5, where the graph tracks the value of inventory $V(t)$ over time. For the purpose of illustration, suppose the jobs are sequenced in numerical order. Then the inventory value is $\Sigma w_i$ for a time interval equal to the processing time of the first job. At that point, the inventory value drops by $w_1$ and remains at that level until the second job completes. The graph follows a stair-step pattern, but in contrast to Exhibit 1.4, the step sizes are not all the same. Each step down corresponds to a reduction in inventory value by one of the $w_i$ values. Again, the objective is the minimize the area under the graph of $V(t)$, and again, we can show that this objective is achieved by the SWPT sequence.

## Shortest-First and Group Technology

An important principle in organizing systems with many elements is to group like elements together. Elaborating on this idea, the group technology principle has been applied effectively in discrete-parts manufacturing. For example, parts can be grouped in to similar families according shapes, sizes, materials, etc. In many manufacturing applications, a machine switches from making products of one family to making products of another family by means of an expensive setup. To avoid excessive setup time, and to simplify scheduling, many production schedules allow just one setup for all the parts in each product family. The jobs in the family are then grouped together so that they can all be done after the same setup. How does the shortest-first principle apply to a sequence of groups?

In the so-called group technology problem, the given information consists of a set of jobs, the family each belongs to, their processing times, and their weights. For family *f*, let $T_f$ represent the sum of the processing times of all the jobs in the family, let $W_f$ represent the sum of the weights, and let $s_f$ represent the length of the setup time required to begin production of the family. Then, to solve the problem of minimizing Σ*wC*, the first step is to sequence the jobs by SWPT within families. Then, to order the families, choose according to the smallest ratio $(s_f + T_f)/W_f$. In effect, this prescription implements SWPT twice. At the level of jobs, the sequencing rule holds, in its usual form. At the level of families, the sequencing rule applies to family processing time (including setup), as scaled by total family weight.

## Limitations of APIs

We have seen that the shortest-first rule yields an optimal sequence when we are interested in simple measures of delay. That is, if the objective is the sum of completion times or the sum of wait times, we know that shortest-first priority is optimal. We also know that this principle generalizes to importance weights if we scale the processing times. Furthermore, an adaptation of the shortest-first rule applies to jobs when they are grouped in families. We might wonder when the shortest-first rule would fail.

As an illustration of the limits of SPT sequencing, consider a situation in which jobs have due dates but no weights. In other words, the given information consists of *n*, the number of jobs, and, for each job, a processing time $t_i$ and a due date $d_i$. The presence of due dates gives rise to a scheduling goal that is different from minimizing delay times or inventory costs. When due dates exist, a reasonable goal is completing jobs by their due dates. One way to quantify this goal is to calculate a *tardiness* value for each job in the schedule. Tardiness is defined as $T_i = \max\{0, C_i - d_i\}$. In other words, tardiness is zero as long as job *i* completes by its due date; otherwise, tardiness is equal to the amount of time that the job is late. A reasonable objective is to minimize the sum of tardiness values in a sequence, for which the usual shorthand is Σ*T*. If we can work to this objective, then whenever there is a sequence that completes all jobs on time, we could produce it by using Σ*T* as an objective and finding its minimum value.

However, finding the minimum value of Σ*T* is not a straightforward problem. To suggest why this is so, let us return to the method of adjacent pairwise interchanges. Specifically, we consider swapping jobs *j* and *k* as in Exhibit 1.3, but now with the Σ*T* objective in mind. In addition, take job *k* to be shorter than job *j*, so that $t_k < t_j$. Consider the situation depicted in Exhibit 1.6, where job *j* is originally on time and *k* is late in schedule S. After the swap, job *k* is on time and *j* is late. However, the increase in the tardiness of job *j* (rightward arrow) is greater than the decrease in the tardiness of job *k* (leftward arrow), as the example is constructed, so that the swap actually makes the total tardiness worse.

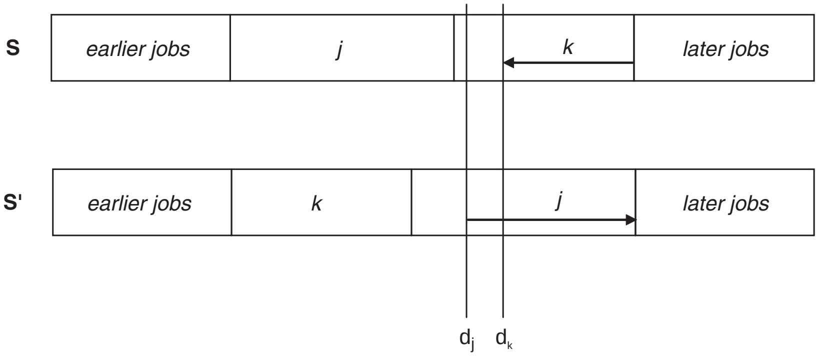


*Exhibit 1.6*

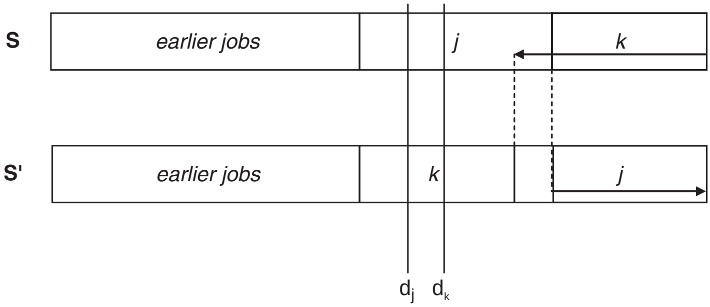


*Exhibit 1.7*

On the other hand, imagine that the pairwise interchange opportunity occurred later in the sequence, as depicted in Exhibit 1.7. In this case, both jobs are late in the initial sequence. However, the swap increases the tardiness of job *j* less than it decreases the tardiness of job *k*, so the net effect is an improvement in total tardiness. (The increase for job *j* is represented by the length of the rightward arrow in Exhibit 1.6, while the decrease for job *k* is represented by the length of the leftward arrow.)

The implication of the cases shown in Exhibits 1.6 and 1.7 is that it is not always a good idea to sequence the jobs in shortest-first order. Moreover, in the case of the Σ*T* objective, the desirability of following the shortest-first rule may hold early in the sequence but not late in the sequence. This was not the case when the objective was the sum of completion times, where the preference for shortest-first sequencing was universal.

Adjacent pairwise interchanges allow us to derive a “delta” measure to account for the relative change in the objective when a job swap is made. Our first such example was the calculation of $\Delta = \Sigma C(\mathrm{S}) - \Sigma C(\mathrm{S}')$. If this calculation leads to a universal result (i.e., “the shorter job should come first”) then we have been successful at determining the nature of the optimal sequence. On the other hand, if this calculation leads to a conditional result, as with the tardiness objective, then the delta measure is not strong enough to dictate the full sequence. In that case, other methods must be brought to bear on the sequencing problem.

## Applications

The sequencing problem arises in many settings. Carrying out design projects, as in our first example, is analogous to writing a set of term papers for a set of course assignments, or producing a set of grant proposals, or drafting a set of legal documents. Whenever there is a collection of tasks waiting for service and a one-at-a-time facility for providing that service, the sequencing problem represents the essential decision. In these examples, of course, the person who makes the sequencing decision often represents the facility providing the service as well, but we could easily separate those two roles.

In manufacturing systems, there is frequently a need to sequence a set of available production tasks that are waiting for a piece of machinery. Often, there is one critical piece of machinery, such as an oven, or there is a "bottleneck" operation, such as inspection or testing, whose schedule strongly influences overall performance. Sequencing jobs at the bottleneck sometimes lends itself to the sequencing analysis of the type we've illustrated in the foregoing sections. Sequencing jobs at an oven may lead to a variation in which several jobs can be processed simultaneously.

Nevertheless, the more typical manufacturing system is highly complex and characterized by shifting bottleneck patterns and unpredictable events. Sequencing models would seem to be overmatched when it comes to scheduling these kinds of systems. However, a generation of research in this area has produced a remarkable insight: simple sequencing initiatives can be adapted very effectively to complex situations. The archetype for such systems is the *job shop* model, in which several different kinds of machines work on a large number of multi-operation jobs. Each job has a unique routing through the machines, and as a job proceeds along this routing, it finds itself waiting for service along with other jobs at the various machines. The vocabulary of "jobs" and "machines" describes some of the traditional settings for a job shop, such as the manufacture of metal parts or electronic components. However, the model may be equally as applicable to the flow of paperwork in the back office of an insurance company or a bank.

As yet, there is no comprehensive theory that prescribes how work should be scheduled in a job shop system. One effective approach is to decompose a complicated, multi-machine system into a collection of one-machine subsystems, where it is easy to implement SPT as a scheduling procedure. Research has demonstrated that if decisions are made locally—that is, if we isolate each subsystem from the others and apply SPT with only local considerations—performance in the overall system can be enhanced. A simple principle such as SPT sequencing, applied locally at each machine or service center, can thus provide major benefits in performance.

## Historical Background

Modern research into sequencing and scheduling problems is usually traced to journal articles of the 1950s. The optimality of shortest-first sequencing for the sum-of-completion-times objective was one of the early results that attracted researchers to the formal study of sequencing and scheduling. Based on its first appearance in an archival journal, the optimality of SPT is usually credited to Smith (1956). A complementary result, published around the same time, showed that sequencing by the earliest due date minimizes the maximum tardiness in the schedule. Soon after, researchers learned that the problem of minimizing the total tardiness in the schedule ($\Sigma T$) was quite challenging. Then research began to generalize the basic results in several directions.

One direction that attracted researchers was elaborating on the elements of the processing model. What happens when there are two machines instead of one? What if they operate in parallel or in series? What if there are three machines? Ten? When there are several machines in the model, it becomes necessary to describe the workflow or *routing*. What if all workflow is the same? What if it is different? What if it is balanced across machines or unbalanced? What if there are assembly operations? What if jobs require both labor and machine resources simultaneously? What if there are breakdowns?

Another direction was an elaboration of job features. What if jobs are not all available at the outset but instead arrive continually, over time? What if jobs arrive randomly? What if work times themselves are also uncertain (not known in advance)? What if work times can be estimated, but with some inherent forecast error? What if there are restrictions on the ordering of certain pairs of jobs?

Yet another direction was elaborating on the basic objectives. The $\Sigma C$ objective begins to capture the concern with inventory or responsiveness, just as the $\Sigma T$ objective begins to capture the concern with meeting due dates, but are there other reasonable ways of quantifying these considerations? What about efficiency? What about costs and profits? What if the manufacturing facility is only a captive resource, working to support a supply chain?

It has not been possible to answer all of these questions in one place, of course, because our knowledge has been developing gradually as sequencing and scheduling problems have been addressed by researchers. Guideposts for the expansion of knowledge have usually appeared in the form of textbooks or proceedings, and occasionally in the form of influential articles. The first major book on the topic was *Industrial Scheduling*, a collection of papers edited by Muth and Thompson (1963) and largely based on the presentations made at a specialized conference two years earlier. The collection reviewed some of the early work from the 1950s and highlighted some of the major research in progress.

The next book in the field was *Theory of Scheduling*, the pioneering textbook written by Conway et al. (1967). This book brought together mathematical theory, queueing analysis, and simulation results, summarizing the major advances and providing new frameworks and concepts for scheduling problems. The text elaborated on the simulation results in Conway (1965a,b), two articles that set the research agenda for simulation approaches to the job shop problem.

The most enduring textbook has been *Introduction to Sequencing and Scheduling*, introduced by Baker (1974). This book focused on deterministic models in sequencing and scheduling and for its time, represented comprehensive coverage of that portion of the field. In its updated form, *Elements of Sequencing and Scheduling* (Baker 2005), the book remains in use as a contemporary text. Less encyclopedic than it was thirty years ago, it nevertheless provides an introduction to basic scheduling concepts and results, and it serves as a stepping stone to the vast research literature on the subject.

Echoing the impact made by the specialized conference that gave rise to the field of scheduling in the early 1960s, two more symposia reinforced the growth of the field and motivated further work. The first symposium led to a collection of papers edited by Elmaghraby (1973), and the second formed the basis for a special issue of the journal *Operations Research* under the editorship of Nemhauser (1978).

The 1970s saw the development in computer science of a body of knowledge relating to computational complexity. Although complexity is a broader topic, many of its early applications were made to problems in scheduling. Furthermore, knowledge in sequencing and scheduling could be organized in a new and informative way with the help of complexity theory. The linkages between classical scheduling and computer science were first highlighted in *Computer and Job/Shop Scheduling Theory*, edited by Coffman (1976). The next major text was *Sequencing and Scheduling: An Introduction to the Mathematics of the Job-Shop*, by French (1982), which drew on developments in complexity theory to organize its coverage of deterministic scheduling. However, the most influential work in this period was perhaps a survey paper by Lawler et al. (1982), which updated the field and created a comprehensive classification of scheduling results to guide the next generation of researchers.

With the field in a more mature state, progress in finding optimal solutions to well-specified scheduling problems began to slow. More attention was paid to heuristic methods. The viewpoint of heuristic thinking was explored in depth in *Heuristic Scheduling Systems*, by Morton and Pentico (1993). They identified the need to integrate operations research and artificial intelligence approaches with sophisticated information systems. Using that philosophy, they described the development of heuristic methods for complex scheduling problems.

A new perspective on scheduling emerged from the artificial intelligence community, exemplified by the collection of papers entitled *Intelligent Scheduling*, edited by Zweben and Fox (1994). Meanwhile, there was considerable progress in the analysis of parallel-machine scheduling, which had elaborated on the computer science perspective that had developed some 20 years earlier. As new contributions were made to those models, updated coverage of the topic finally appeared in the book *Scheduling Computer and Manufacturing Processes*, edited by Blazewicz (1996).

The latest new book on sequencing and scheduling is *Scheduling: Theory, Algorithms, and Systems*, by Pinedo (2001). It is the first text to integrate stochastic and deterministic scheduling models and appears to be the leading textbook on scheduling currently on the market.

## Selected Bibliography

Baker, K.R. (1974) *Introduction to Sequencing and Scheduling*, John Wiley, New York.

Baker, K.R. (2005) *Elements of Sequencing and Scheduling*, Tuck School of Business, Hanover, NH.

Blazewicz, J. (1996) *Scheduling Computer and Manufacturing Processes*, Springer-Verlag, Berlin.

Coffman, E.G. (1976) *Computer and Job/Shop Scheduling Theory,* John Wiley, New York.

Conway, R.W. (1965a) "Priority Dispatching and Work-in-Process Inventory in a Job Shop," *Journal of Industrial Engineering* 16, 123–130.

Conway, R.W. (1965b) "Priority Dispatching and Job Lateness in a Job Shop," *Journal of Industrial Engineering* 16, 228–237.

Conway, R.W., W.L. Maxwell, and L.W. Miller (1967) *Theory of Scheduling*, Addison-Wesley, Reading, MA.

Elmaghraby, S.E. (1973) *Symposium on the Theory of Scheduling and its Applications*, in *Lecture Notes in Economics and Mathematical Systems*, Springer-Verlag, Berlin.

French, S. (1982) *Sequencing and Scheduling: An Introduction to the Mathematics of the Job-Shop,* Ellis Horwood, West Sussex (distributed by John Wiley).

Lawler, E.L., J.K. Lenstra, and A.H.G. Rinnooy Kan (1982) "Recent Developments in Deterministic Sequencing and Scheduling: A Survey," in M.A.H. Dempster, J.K. Lenstra and A.H.G. Rinnooy Kan (eds.) *Deterministic and Stochastic Scheduling*, D. Reidel Publishing, Dortrecht, 25–73.

Morton, T.E. and D.W. Pentico (1993) *Heuristic Scheduling Systems*, Prentice-Hall, Englewood Cliffs, NJ.

Muth, J.F. and G.L. Thompson (1963) *Industrial Scheduling*, Prentice-Hall, Englewood Cliffs, NJ.

Nemhauser, G.L., (ed.) (1978) *Scheduling*, a special issue of *Operations Research* 26, 1.

Pinedo, M. (2001) *Scheduling: Theory, Algorithms, and Systems*, Prentice-Hall, Englewood Cliffs, NJ.

Smith, W.E. (1956) "Various Optimizers for Single-Stage Production," *Naval Research Logistics Quarterly* 3, 59–66.

Zweben, M. and M. Fox (1994) *Intelligent Scheduling*, Morgan Kaufman, San Francisco.

# Chapter 2
# The Knapsack Problem

**John J. Bartholdi, III**
**Georgia Institute of Technology**

*The "knapsack problem" appears in many forms in economics, engineering, and business: any place where one must allocate a single scarce resource among multiple contenders for that resource. It has acquired the fanciful name "knapsack problem" because our common experience of packing luggage expresses something of the flavor of the problem: What should be chosen when space is limited?*

## Introduction

Alice and Bob were excited about the bicycle tour they had long planned. They were going to ride during the day, carrying only light supplies, and stay in hotels at night. Alice had suggested they coordinate packing to avoid duplication and extra weight.

Alice was to pack tools: a compass, spoke wrench, chain breaker, hub wrench, and spare tire tube; and Bob was to pack consumables: Water, toilet paper, sunscreen, trail mix, soap. They agreed that neither would pack more than $c = 7$ pounds of supplies.

On the night before the trip Alice gathered everything on the kitchen table and prepared to pack; but it quickly became clear that she had staged rather more than 7 pounds. Decision time.

Alice was a perfectionist. She realized the importance of packing the right stuff and so was prepared to think carefully about it. She methodically indexed the candidate items and listed the weight $w_i$ of each item $i$ (Table 2.1, 2nd row).

But what to pack? Alice saw that she could fill the knapsack to the limit with items 1, 2, and 4. But then again she could just as well fill it with items 3 and 5 or with items 2, 4, and 5. Which is best? Or might it be better to pack only items 2 and 3 even if this alternative leaves some residual capacity unused? Why not pack only items 1 and 5 or only 3 and 4?

Alice realized that each item would either be packed or left behind—2 choices, and so there were up to $2^5 = 32$ possible ways to pack the knapsack. Not all of these are possible or even desirable, but it seemed as if she would have to consider each possibility explicitly and choose the best, whatever that meant.

Alice saw that, to choose the best, she needed some way to distinguish the "value" of one packing from another. Was it better to pack the spoke wrench or the spare inner tube?

D. Chhajed and T.J. Lowe (eds.) *Building Intuition*: *Insights From Basic Operations Management Models and Principles*.
doi:10.1007/978-0-387-73699-0, © Springer Science + Business Media, LLC 2008

**Table 2.1** Weight in Pounds of Each Item to be Packed

| Item | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Weight (lbs) | 3 | 2 | 4 | 2 | 3 |

**Table 2.2** Relative Values of Each Item to be Packed

| Item | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Value | 10 | 6 | 11 | 4 | 5 |

She thought she could answer simple questions like this, which asked her to compare the values of individual items, but it seemed harder to compare entire packings.

A good rule of thumb in making hard decisions is to formalize the choices in some way, in hopes of clarifying them. Alice did this by assigning “values” $v_i$ to each candidate item $i$ (Table 2.2, 2nd row).

Furthermore, it seemed reasonable to measure the value of a packing by the sum of the values of the items packed. Under this assumption it became clear that the packing of items 2, 4, and 5, which has a total value of 6 + 4 + 5 = 15, is less desirable, even though it fills the knapsack, than the packing of items 2 and 3, which does not fill the knapsack but has total value 6 + 11 = 17.

Alice saw that she could search for the best items to pack by the following process. Write out all 32 possible subsets of items, sum the values of the items in each subset, eliminate those that are not feasible (exceed the weight limit), and choose the subset of largest value. Tedious perhaps, but simple, straightforward, and guaranteed to produce a packing of maximum value.

Of course, this instance of a knapsack problem is so small that it presents scarcely any challenge. If Alice had had to choose from among, say, 20 items; then there would have been $2^{20}$ = 1,048,576 possibilities and Alice would likely have been awake all night to evaluate them all. Things can become much worse quickly for larger versions of this problem: To choose from among 100 items requires evaluating $2^{100}$ = 1,267,650,600,228,229,401,496,703,205,376 possibilities and it is not likely Alice could evaluate all of them within her lifetime.

## *General Structure of a Knapsack Problem*

In her professional life Alice is a fund manager who oversees investment portfolios. She is currently considering over 100 potential investments and has estimated the return expected from each one. But each investment has a certain cost and Alice may not exceed her budget. Which investments should she choose? You will recognize this as having the same structure as Alice’s problem of packing for the bicycle trip: How to choose a subset of items of maximum value while not exceeding a limit on total “weight.” For the bicycle trip, “value” was a subjective expression of importance and “weight” was physical weight. For the investment portfolio, “value” is estimated return, perhaps constructed via a financial or economic model; and “weight” is the

**Table 2.3** Bang-for-Buck of Each Item to be Packed

| Item | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Bang-for-buck | 10/3 | 6/2 | 11/4 | 4/2 | 5/3 |

cost of investment. The common structure of these problems of choice is called the “knapsack problem” and it is distinctive in the following ways.

- We are faced with a number of yes/no decisions.
- The decisions are independent, except that each yes decision consumes some known amount of a single, common, scarce resource.
- None of the data will change during the time we make our decisions.

The most important part of this problem is the single, scarce resource. We are challenged to use it wisely to achieve the greatest value. For the bicycle trip, that resource was weight capacity; for the investment portfolio, that resource is budget. The knapsack problem is a way of looking at decisions that focuses on that single constraint.

In real life there are not many decisions with single constraints; but there are many in which one constraint matters more than others, and so the knapsack model is much more widely applicable than the idea of a single constraint might suggest. For Alice and Bob there may be an additional constraint in the physical volume that can be packed in the knapsack, but they may be justified in ignoring this constraint for now, perhaps because from previous experience they expect to reach the 7-pound limit well before filling the knapsack. Other constraints might not have well defined thresholds. For example, the cost of supplies may be a “soft” constraint: As more items are packed, Alice and Bob might feel increasingly uncomfortable at the thought of all they must purchase, but there is no clear threshold as there is for the capacity of the knapsack. In such cases, one identifies the most important constraint and models that are in the knapsack formulation. Then, after solving, check that the remaining, unexpressed constraints are satisfied.

For the bicycle trip, Alice and Bob have decided that weight is the most significant issue to them and they have stipulated a limit on it. After finding a good packing, Alice may check the resultant volume to see whether it is acceptable. If not, she might re-pose her problem as one of packing the most valuable load subject to a limit on the total volume. This would be a knapsack problem as well, but with volume as the scarce resource.

The knapsack problem Alice faces as a fund manager is too large to solve by considering every possible solution (a process known as “total enumeration”); but the economic context gave her some insight that helped simplify her decision-making. In the business world, one hopes to achieve a high return on investment (ROI), because this indicates efficient use of financial resources. Accordingly, Alice prefers an investment that promises a greater return per dollar invested (“bang-for-buck”). For the bicycle trip, analogous reasoning suggests that she should prefer to pack items that have a large value-per-pound ratio $v_i/w_i$, Table 2.3.

Alice realized she could avoid the work of total enumeration by simply choosing those items with greatest bang-for-buck. More formally, we can summarize the decision process as follows.

Step 1: For each item compute its bang-for-buck $v_i/w_i$.
Step 2: Sort the items from greatest to least bang-for-buck.
Step 3: Choose items from the top of the list until the next item (call it item $k$) does not fit.

Using this procedure to pack for the trip, Alice would choose items 1 and 2, which together weigh 3 + 2 = 5 pounds and are of value 10 + 6 = 16. We know this is not the best possible solution—for such a small problem you can surely see better packings—but it was very easy to generate. Moreover, we have some reason to believe the solution might not be too far from the best because there is an undeniable logic behind the procedure: Choose those items that best use the scarce resource.

## Bob Packs

Bob was to pack the consumables, which, for pedagogical convenience, had exactly the same weights as given in Table 2.1. He noticed immediately that he had more choices in packing than did Alice: Because his items were divisible, he could repackage any of them and take only a portion; therefore his problem was to decide not which to pack, all or nothing, but the fraction of each item to pack. Like Alice, he saw that some items were more important than others. For example, he valued water more than soap, which they could do without if necessary. Bob listed the relative values of each item, which, again for pedagogical convenience, were identical to those of Table 2.2.

**Building Intuition** The procedure defined by steps 1, 2, and 3 is of a type known as “greedy” because it simply sorts possible choices by some measure of attractiveness (in this case, bang-for-buck) and chooses from the top. It never reconsiders decisions. However, since this procedure is not guaranteed to find the best answer, we call it a “heuristic.” Hence choosing items by bang-for-buck is a “greedy heuristic.” Greedy procedures to find solutions to decision problems are appealing since, in general, they are intuitive.

Although a greedy procedure does not always optimally solve the knapsack problem, there are instances of decision problems in operations management where such a procedure does find the best solution. One such example is the weighted completion time problem discussed in Chap. 1. As is shown in that chapter, a sequence (ordering) of tasks that minimizes the sum of weighted completion times is found by computing, for each task, the ratio of task time divided by weight and then sequencing the tasks in order of nondecreasing ratio values.

Now Bob's packing problem was similar to Alice's: How much of each item to pack so as to maximize the total value of a load that may not exceed 7 pounds.

## *Bob's Solution*

Bob tried Alice's suggestion of choosing the items of greatest bang-for-buck; but when he had chosen items 1 and 2, he noticed that they weighed only 3 + 2 = 5 pounds, which left unused a capacity for 7 − 5 = 2 pounds. Because his items were divisible, Bob decided to go one step further and pack 2 pounds of item 3, which was the next-most-attractive. Two pounds of item 3 represented half its weight and so may be assumed to contribute half the value, or 11/2 = 5.5. This results in a packing that uses all available capacity and achieves a value of 21.5. Furthermore, as can be verified by inspection of this small example, this is the very best packing achievable. It provides the best packing possible in all instances because every single unit of the scarce resource is devoted to that item, or fraction thereof, that returns the greatest possible bang-for-buck.

Let us make Bob's decision process more specific, because, if we can formalize it, then we can program a computer to do it for us.

Step 1: For each item $i$, compute its bang-for-buck $v_i/w_i$.

Step 2: Sort the items from greatest to least bang-for-buck.

Step 3a: Choose items from the top of the list until the next item (call it item $k$) does not fit.

Step 3b: Take only as much of item $k$ as will fill the remainder of the knapsack.

Note that, for any capacity and for any set of items selected in this manner, Bob will have to repackage at most a single item, the one selected in less than full quantity.

For this version of the knapsack problem, in which items are continuously divisible, the procedure is optimal: There is no alternative choice of items to pack that will have greater total value. Let us call the resultant value $V^*$. It is the largest possible value that can be realized from these candidate items subject to this weight limit.

## Alice Tries Harder

Alice was pleased to see that choosing by bang-for-buck always generated the very best possible choices for Bob; but she saw that it would not work for her because her items were indivisible. Half a chain breaker would be useless: Each of her items had either to be packed in its entirety or else not at all and so she could not "top off" remaining capacity by adding just a little of another item.

## *Alice Finds a Guarantee*

Alice saw that her procedure was similar to Bob's except that it stopped one step short[1]. (Step 3b). Surely her solution must be almost as good as Bob's, and Bob's was the best that could be hoped for. Surely, the value of the items greedily chosen could never be "too far" from *V**, the very best that would be achievable if Alice's items were divisible. Alice reasoned thusly: Suppose the heuristic halts, unable to pack item *k* entirely in the remaining capacity of the knapsack. If item k were divisible, so that I could pack it to fill the remaining capacity, it would add value $(v_k/w_k)$ × (the remaining available capacity of the knapsack) to what had been already packed and this would be the best possible packing. But my items are not divisible and so this solution is an ideal, possibly unachievable. If I stop here, omitting item *k*, then the value of what I have packed must be within

$$(v_k/w_k) \times \text{(the remaining available capacity of the knapsack)} \tag{1}$$

of *V**.

In short, Alice cannot have missed optimal by more than the value of that *k*-th item, the first that that did not fit.

## *How Wrong Can Alice Be?*

Expression (1) bounds the opportunity Alice might forfeit by accepting an easy-to-compute solution rather than insisting on an optimal solution. Here is an example that embarrasses the greedy heuristic: Consider a knapsack of capacity *c* for which two items contend. One is of weight 1 and value 1; the other is of weight *c* and value just shy of *c*, call it $c$-ε. The greedy heuristic would choose the first item and halt, because the second item no longer fits. The total value achieved would be 1; yet if the second item had been chosen the total value would have been $c$-ε. Since *c* could be any large number, the error, both relative and absolute could be arbitrarily large.

That is the worst case. What might we actually expect? In some circumstances this worst-case error is small, perhaps small enough that we would feel comfortable accepting it. For example, we would expect the worst-case error (1) to be small and the solution to be quite close to *V** whenever any of the terms of (1) are small; that is, in any of the following situations.

- If $v_k$ is small compared to the total value packed; or
- If $v_k/w_k$ is small compared to the average bang-for-buck of the packed items; or
- If the remaining capacity of the knapsack is a small fraction of the total.

[1] Alice's procedure could be made a little more effective by having it try to fit each successive candidate item into the knapsack rather than stopping at the first failure. This will not affect our subsequent discussion.

All of these conditions, and especially the third one, may be expected to hold if most items are small with respect to the capacity of the knapsack. In such case we are packing lots of small items into a large container, which, as we know from experience, is easy to do well (think socks in a suitcase). Many items will fit before the first one fails, and consequently that item may be expected to have a relatively small bang-for-buck $v_k/w_k$. This is so for the simple reason that we will have already packed those items that have the greatest bang-for-buck. Moreover, because the items are small with respect to the capacity of the knapsack, we would expect any capacity unused after Step 3 to be small. Consequently we would expect the solution to be quite close to $V^*$, the upper bound. For example, if no item weighs more than x% of the weight limit, then we can be sure that a greedy packing will always fill the knapsack to within x% of its capacity and provide total value within x% of the maximum possible. Statistics are on our side here.

## *Is It Good Enough?*

Is it "good enough" to know that your solution is close to the best possible? The answer to this question is the ubiquitous "it depends." In particular, it depends on how much you are willing to pay in money, time, and effort to get a better solution. It is probably not worth worrying over if you are packing a knapsack for a bike trip; but it may well be worth anguish if you are loading a space vehicle for an expedition to Mars.

There are some situations in which it makes sense to accept an approximate solution, even if you wish for the best. For example, the best may not be well defined when the data are uncertain. For the bike trip, the weights of the items to be packed can easily be measured and everyone can agree on them, assuming we have a scale of sufficient accuracy. It is not so clear how to measure value. It may be that Alice and Bob disagree on the value of sunscreen; indeed, Alice might be uncertain herself exactly what number to assign to the value of a compass. When the data are uncertain, the expense of careful optimization may not be worthwhile.

## Alice Is Difficult to Please

Alice was a perfectionist. She took pride in owning a top-of-the-line touring bicycle, which had been fine-tuned for lightness and strength. She was dissatisfied that the greedy solution to her knapsack problem might not be the best possible. She wanted to know for sure whether it was the best; and if it was not … well, then she wanted the best. As we saw, Alice could be sure by methodically evaluating all possible solutions and choosing the best. But this is practical only when the number $n$ of candidate items is quite small because the work to evaluate $2^n$ possibilities increases very rapidly in $n$ and so is impractical for any but the smallest instances of the knapsack problem.

It is possible to compute an optimal solution for fairly large problems by trying carefully to fit items into the knapsack, adding one after another, possibly removing some if we have come to an impasse, and resuming. We can formalize such a process by representing it as a network that summarizes the sequence of decisions about what to pack, the resultant state of the partially packed knapsack, and the remaining choices available to us. For example, Fig. 2.1 shows the network of decisions representing Alice's problem if she, unlike Bob, could not split any item.

There are $1 + n = 6$ columns of vertices indexed by $i = 0, 1,\ldots,n$; the first vertex corresponds to the start (an empty knapsack) and each of the next 5 columns corresponds to the disposition of an item. Each column is composed of $1 + c = 8$ rows indexed by $w = 0, 1,\ldots,c$, each corresponding to one additional unit of weight. Each vertex $(i,w)$ represents a possible state of a partially packed knapsack, the state in which items $1,\ldots,i\text{-}1$ have been considered and either packed or rejected, resulting in a knapsack of weight $w$. Each edge is directed from left to right, bottom to top, and represents the inclusion or exclusion of an item. In particular:

- Any edge from $(i\text{-}1,w)$ to $(i,w)$ represents the exclusion of item $i$: We have considered item $i$ but not packed it and so the cumulative weight of the partially filled knapsack remains $w$. Such an edge is assigned length 0.

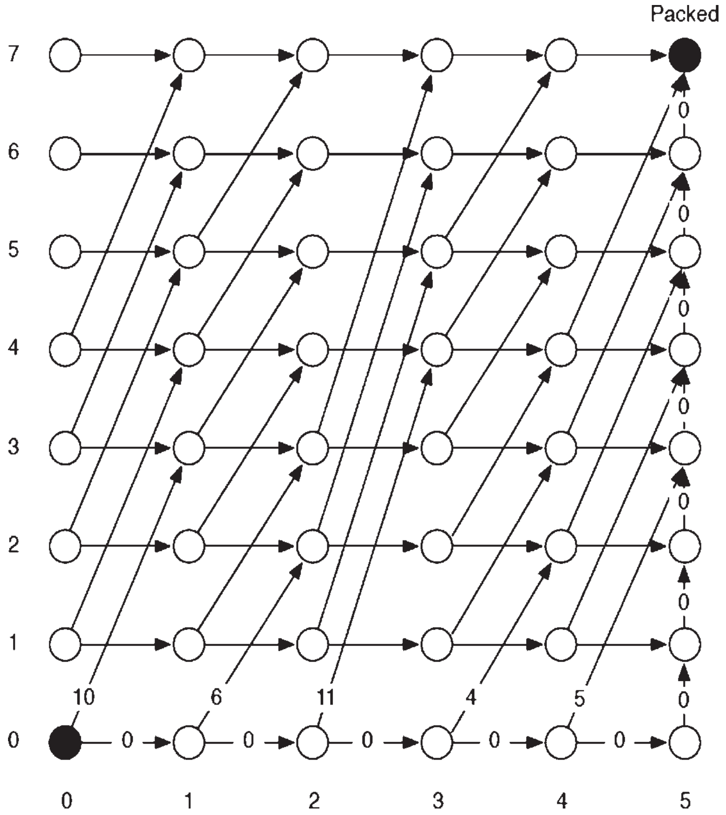


**Fig. 2.1** Alice's decision modeled as a problem of finding the longest path through a network

- Any edge from *(i-1,w)* to *(i,w + $w_i$)* represents the inclusion of item *i*: We have considered item *i* and packed it and so the cumulative weight of the knapsack has increased by $w_i$. Such an edge is assigned length $v_i$.
- Any edge from *(n,w)* to *(n,w + 1)* represents the decision to leave one unit of capacity of the knapsack unused. Such an edge is assigned length 0.

NOTE: To avoid clutter in the figure, only a very few edges have lengths indicated on them. The complete figure would have a length indicated on each edge.

Here is the key insight: Any path from the bottom-left origin *(0,0)* to the vertex on the top right *(n,w)* corresponds to a sequence of decisions of what to put in the knapsack and what to omit. We want a selection of items that gives greatest value, which means we want the longest-path from *(0,0)* to *(n,c)*.

Fortunately, there are simple methods to compute the longest path in a network such as this, in which all edges point in the same direction (so that there are no cycles). The standard method is called dynamic programming and it works like this:

- Start at the last column and label each vertex in this column with 0, which is the length of the longest path from it to vertex *(n,c)*.
- Working backward to preceding columns, label each vertex in the current column with the largest value of: the length of an edge departing this vertex plus the label of the vertex on the other end of that edge. This will be the length of the longest path from the current vertex to vertex *(n,c)*.

When we have finished labeling the first column, the label of vertex *(0,0)* will give the length of the longest path to vertex *(n,c)*, which will also, by the clever way we have built the network, be the maximum value of any way of packing the knapsack. The path of this length, which we can discern by the pattern of labeling, tells us exactly which items to pack and which to leave, Fig. 2.2.

As you will have surmised, this method is tedious for a human but simple for a computer. One can think of it as merely a clever way of organizing a search. It reduces the work by taking advantage of the observation that the best-packed knapsack of capacity *c* must also contain the same loads as the best-packed knapsacks of smaller capacity. In any event, this methodical search takes substantially less time than evaluating every one of the $2^n$ possible loads.

## *The Cost of Finding Optimum*

To find the longest path on the network above requires us to examine about *nc* vertices[2], and so we may take this as an estimate of the total work required. Notice that this depends on the number of items to be packed, which seems reasonable; but it also depends on the capacity *c* of the knapsack. This is more troubling because it seems

[2] Some vertices are clearly spurious to the solution, such as those in the upper left-hand corner, but it is generally not worth the effort to identify those that can be ignored.

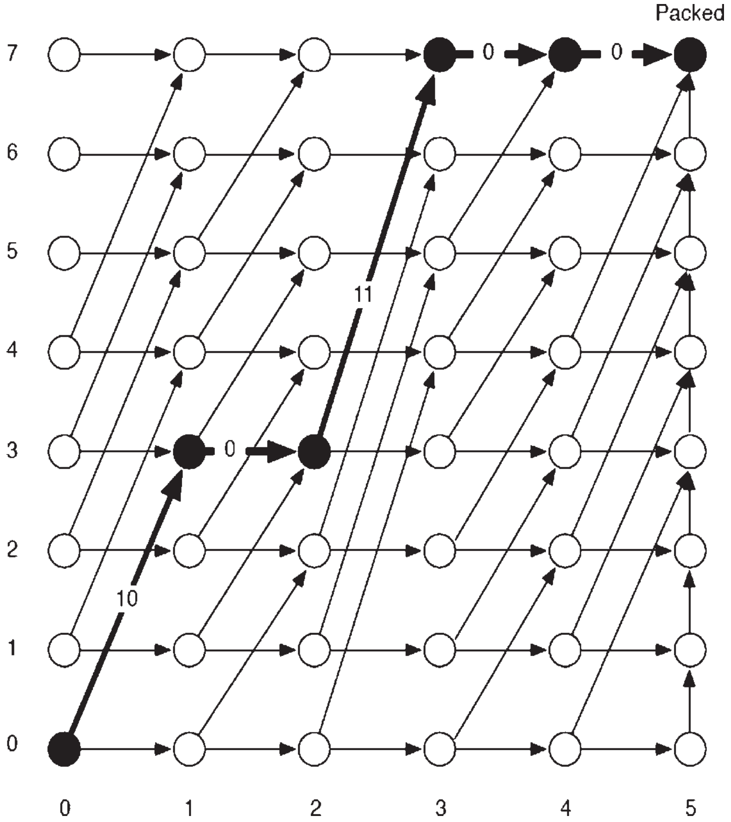


**Fig. 2.2** The longest path corresponds to packing items 1 and 3 and omitting the remainder. The total value is 21, the length of the path

to suggest that the work can depend on how accurately we measure the capacity. For example, the network to represent Alice’s problem would have been larger if she measured all weights to the nearest ounce, and larger still if she measured to the nearest tenth of an ounce. This observation works in the other direction as well. Alice could reduce the size of the network, and presumably the work of computing a solution, by measuring less accurately: to the nearest kilogram instead of to the nearest pound. This reveals an interesting feature of the knapsack problem: The greatest determinant of the work to construct an optimal solution is the precision of the weights $w_i$ more so than the number of items $n$. The more precise the data, the more work is required to take advantage of that precision. The work increases exponentially with any increase in precision but only linearly in the number of items. For example, if capacity of the knapsack had been specified as 7.01 pounds instead of 7 then the network would require 701 rows of vertices and the work to solve would have increased by a factor of 100.

## Pedaling Into the Sunset

As Alice and Bob have determined, we can approach any problem with knapsack structure via the most appropriate of the following:

- If there are few items from which to select, we can simply enumerate all possibilities and choose the most valuable subset of items.
- If there are a moderate number of items from which to select or if the precision of the weights is not too great then we can compute the optimal solution by dynamic programming, as on the network above.
- If there are very many items contending for selection or if the data are very precise (so that it is impractical to compute an optimal solution) or if the data are very imprecise (so that an optimum solution is not to be relied upon), then the greedy heuristic—choosing items of greatest bang-for-buck—is suitable.

In many practical applications, there is much merit in Alice's greedy heuristic if the potential shortfall in quality of solution can be tolerated. Besides ease of computation, Alice's solution has the advantage that it is simply a sorted list of all the items, ranked from greatest to least bang-for-buck. If the bang-for-buck of an item were to change, it is a simple matter to move this item to its new position in the sorted list. This allows us to incrementally adjust the knapsack solution as data changes, so that the greedy solution can support *dynamic decision-making*. That is, we might monitor items and remove or insert them as their values or weights change or as the capacity of the knapsack changes. This would be useful, for example, in maintaining an investment portfolio that must adapt to changes in the financial marketplace, or in keeping the right product stored in the right locations of a distribution center even as the patterns of customer orders change.

## Applications of the Knapsack Problem

Among the straightforward applications of the knapsack problem are, unsurprisingly, problems of loading shipping containers especially when one of weight or volume is known in advance to be the limiting constraint. These issues can be quite significant when space is scarce, as is the case when manufactured product is shipped to the US from China in advance of the holiday selling season.

Another common application is to cutting stock problems. For example, paper mills produce huge rolls of paper, the dimensions of which are determined by the manufacturing process. These roles are subsequently sliced into smaller rolls to fill customer orders. The value of the smaller rolls depends on the selling price. A similar problem is faced by manufacturers of fiber optic cable, who must decide how to cut lengths of cable to satisfy customer orders while extracting the greatest value from each length of cable.

As suggested earlier, the knapsack problem is a basic tool of portfolio optimization, where budget is almost always the most important constraint. Many portfolio optimization problems generalize the knapsack problem in ways that more exactly model economic phenomena. Thus it may be possible to purchase 0, 1, 2, or more shares of an investment; and the additional shares may bring diminishing returns. Many of the ideas mentioned by Alice and Bob can be extended to account for such additional complexities.

In a warehouse or distribution center (DC), space is frequently a scarce resource because of inevitable growth in the number of products handled. There is competition among the stockkeeping units for storage in the most convenient areas of the DC, where customer orders can be filled most quickly. This can be modeled as a knapsack with space as the limited resource and labor-savings as the value of storing a product in a convenient location.

Many scheduling problems can be posed as knapsack problems with machine time the most important scarce resource. This is especially appropriate when the machine represents a large capital investment. In such case it can make economic sense to schedule the machine, perhaps by solving a knapsack problem, and then purchasing whatever additional resources, such as workforce or transportation or raw materials, as are necessary to meet the schedule.

The knapsack appears as a sub-problem in many, more complex mathematical models of real world problems. One general approach to difficult problems is to identify the most restrictive constraint, ignore the others, solve a knapsack problem, and somehow adjust the solution to satisfy ignored constraints.

## Historical Background

The knapsack problem seems to have first been identified in print in 1957, in two important publications. One was a paper by George Dantzig (1957), one of the developers of linear programming and a creator of the field of Operations Research. He showed that the continuous version of the knapsack problem (the one faced by Bob) is perfectly maximized by choosing items by bang-for-buck.

The problem must have been a topic of conversation amongst the early specialists in discrete optimization because in the same year Richard Bellman, another important early figure in Operations Research, described how to use dynamic programming to solve the knapsack problem. (This is equivalent to the method above in which we find the longest path on a special network.) Very quickly thereafter the knapsack model was applied to a range of applications, including most of those listed above.

Throughout the 1980s there was much work on approximation algorithms, solution techniques that cannot be guaranteed to produce optimal solutions because they strategically give up some quality to achieve speed of execution. The knapsack problem was a popular target for the development of such approximation methods. Indeed, one of the first “polynomial approximation schemes” was developed for the knapsack problem by Sahni (1975). Such schemes may be thought of as solution methods in which one may specify a desired guarantee for the quality of solution in advance of solving. As to be expected, the work to solve increases quickly with the quality required. Immediately afterwards, this was improved by Ibarra and Kim

(1975) to a “fully polynomial approximation scheme,” which makes the trade-off between quality of solution and effort slightly more favorable. Specialized solution techniques to solve the knapsack problem and its variants are provided by Martello and Toth (1990). More recently, researchers, such as Kellerer et al. (2005) have worked on ways to exactly solve ever larger instances of the knapsack problem. Many of these involve solving some “core” of the problem and then building this partial solution to a full solution.

Interestingly, the knapsack problem figured prominently as the first suggested basis for a public key encryption system. This early work is described in Diffie and Helman (1976) and Merkle and Helman (1978). It should be noted that the approach was later “cracked” by cryptographers and replaced by more resistant schemes.

As an example of the flexibility conferred by the simplicity of the greedy algorithm, Bartholdi and Hackman (2006) make extensive use of the knapsack problem as part of a larger model to cache product in a distribution center.

## Selected Bibliography

Bartholdi, J. J. III and S.T. Hackman, (2006). *Warehouse and Distribution Sciences*. Georgia Institute of Technology, Altanta, GA. This book is freely available at www.warehouse-science.com.

Bellman, R. (1957) *Dynamic Programming*, Princeton University Press Princeton, N.J., Reprinted (2003) Dover Publications, Mineola, N.Y.

Dantzig, G. (1957). “Discrete variable extremum problems”, *Operations Research* 5,266–277.

Diffie, W. and M. Helman, (1976) “New directions in cryptography”, IEEE Transactions on Information Theory 22(6):644–654.

Ibarra, O. H. and C. E. Kim, (1975) “Fast Approximation Algorithms for the Knapsack and Sum of Subset Problems”, Journal of ACM 22,463–468.

Kellerer, H., U. Pferschy and D. Pisinger (2005) *Knapsack Problems*, Springer, Berlin, Germany.

Martello, S. and P. Toth, (1990). *Knapsack Problems: Algorithms and Computer Implementation,* John Wiley, Ltd. This book is freely available at www.or.deis.unibo.it/knapsack.html.

Merkle, R. and M. Helman, (1978). “Hiding information and signatures in trapdoor knapsacks”, IEEE Transactions on Information Theory 24(5),525–530.

Sahni, S. (1975). “Approximate algorithms for the 0–1 knapsack problem”, *Journal of ACM* 22:115–124.

# Chapter 3
# Flexibility Principles

**Stephen C. Graves**
**Massachusetts Institute of Technology**

*Consider a setting with multiple demand classes that are served by a set of resources. When each resource is limited to serving only one demand class, we can often have a situation where some resources are under-utilized and idle, while others are over-utilized and not able to meet the demand. One tactic for dealing with this situation is to make each resource more flexible so that it can serve more than one demand class. But how much flexibility should each resource have and what is the best way to deploy flexibility across the resources? This chapter shows that when done right, limited flexibility can provide almost the same level of benefits as complete flexibility.*

## Introduction

FCM is a small manufacturing company that assembles a family of fuel controllers for jet engines and turbines. The assembly of each type of fuel controller is a complex, labor-intensive activity that requires specialized tools and skills and takes half of a work day; that is, the assembly time for each unit is 4 h. FCM supplies six distinct products, labeled A, B, C, D, E, and F, and has six certified technicians, Adam, Bob, Carol, Dianne, Edith, and Fred. Each technician has been trained to assemble one type of fuel controller, as shown is Fig. 3.1.

At the start of each week, FCM receives orders from its customers, which are Original Equipment Manufacturers (OEMs) for which the fuel controller is a subsystem required for their final products. The OEMs expect delivery of the completed orders at the start of the following week. Thus, FCM has exactly 1 week to meet the orders.

FCM follows a simple protocol for running its assembly operation. At the start of each week, it schedules each technician to assemble the number of units exactly equal to the demand for his/her product. For instance, if demand for product A is 8 units, then Adam will assemble 8 units, no more and no less, in the coming week so as to meet the demand for A.

Based on past order history, FCM has developed a forecast for demand for each product. For instance, FCM expects demand for product A to be 10 units on

D. Chhajed and T.J. Lowe (eds.) *Building Intuition*: *Insights From Basic Operations Management Models and Principles*.
doi:10.1007/978-0-387-73699-0, © Springer Science + Business Media, LLC 2008

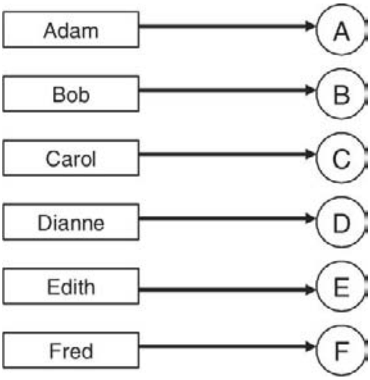


**Fig. 3.1** Base case at FCM, a dedicated system

average. However, this is just a forecast and actual demand for A can deviate from the forecast. Indeed, weekly demand for product A has ranged between 8 and 12 units, and FCM forecasts that the likelihood that demand is 8, 10 or 12 units is equally likely. [Customer orders are typically for pairs as each engine will require two fuel controllers.] That is, the probability is 1/3 for each of the possible outcomes, 8, 10 or 12 units.

Similarly, FCM has made a forecast for demand for the other five products, and each product has the same demand forecast. For each product the possible demand outcomes are 8, 10, or 12, with each occurring with equal likelihood, namely each with probability of 1/3. Furthermore, the demand for each product is independent from week to week and independent of the demand of each of the other products. Thus, the fact that the demand for product A in the current week is 12 has no influence on the demand outcome for product B or on the demand outcome for A in the following week.

This is a quite stylized and simplified example, but is representative of many operational settings in which there are multiple demand categories or classes (products A, B,..F) that are served by a set of resources (Adam, Bob, … Fred). Many service systems operate with parallel servers and different demand classes. For example, the resources or servers might be physical therapists, or barbers, or tutors, and the categories of demand would correspond to customers with differing service requirements (e.g., therapy for different body parts; hair cuts for girls, boys and adults; help with math, science, or history). In a manufacturing operation, we might have a set of equipment that serves different types of jobs. For example, the photolithographic equipment in a semi-conductor fabrication plant is responsible for the two to three dozen photolithographic process steps that are required for each lot of wafers that flow through the facility. Each step requires a different setup and specific tooling, e.g., a mask. As a consequence, a plant will often dedicate each process step to a limited subset of the available equipment.

A challenge in these settings is how best to match the resources to the demand. At FCM we have six technicians, each with a nominal work week of 40 h. Thus, if demand for each product were 10 units each week, then we would have a perfect match between the available resources and the demand. Each unit requires 4 h of assembly time, so each worker would complete his/her 10 units in his/her normal work week of 40 h. Unfortunately, when demand can vary from week to week, we lose this balance. When demand for product A is 12 units, Adam needs to work 48 h, or 6 days, to complete the work within the week; Adam is willing to do this, as he gets paid an overtime rate of 150% of his normal pay for the extra 8 h he works. When demand for product A is 8 units, Adam is underutilized and needs to work only 32 h. However, FCM still pays him for a full week, namely for 40 h, and will assign him other tasks to fill out the remaining 8 h of work time.

Because of this imbalance, each technician works a day of overtime once every 3 weeks on average, and thus, FCM pays, on average, for 2 days of overtime each week. FCM naturally wonders about how to mitigate or avoid this overtime expense. In such settings, the first thought is to explore how the variability might be eliminated or reduced. If FCM might somehow entice the customers to order exactly 10 units of each product each week, then there would be a perfect match of demand and resources and no need for overtime. This might be possible if there were a few large customers, and if FCM were willing to provide an incentive, like a price discount for stable ordering. However, this could be more costly than the overtime, and/or impossible to achieve if there were lots of small customers.

If FCM cannot affect the demand variability, then it might consider the common operational tactics of creating a buffer of some form as a counter measure to the variability.

One form of buffer is a *capacity buffer*; FCM might hire and train a seventh technician, who would be capable of assembling any of the products (e.g., Jack, a jack-of-all trades). Then each week Jack would pitch in and help out with any demand surpluses; that is, he would assemble the 11th and 12th units for any product that has demand of 12. This would eliminate the need for overtime, with one exception. If the demand for each product were 12, then Jack would have 12 units to assemble and he would need to work an extra day of overtime; but this would occur very rarely, with probability $\left(\frac{1}{3}\right)^6 = \frac{1}{729}$, on average once every 729 weeks (about 14 years). Thus, adding capacity in the form of an additional technician effectively eliminates the need for overtime. The cost, though, might be quite high; we now need an extra person, fully trained to assemble every product. In most settings, this cost is likely to exceed the savings, namely saving 2 days of overtime each week, unless we can find other things for Jack to do.

A second way to buffer variability is to create a *time buffer*. Currently FCM commits to deliver orders within a week. By lengthening its delivery lead time from 1 week to, say, 2 or 3 weeks, FCM could create a time buffer that would allow it to smooth out the week to week demand variability. This could substantially reduce FCM's overtime, but only if its customers were willing to accept a longer delivery lead time.

The third type of buffer is to use an *inventory buffer*. FCM could assemble its fuel controllers to stock, and then serve demand from this inventory. Similar to a time buffer, the inventory buffer would allow FCM to smooth its production and reduce the amount of overtime. Making to stock presumes that each product is not customized to an order and that each product has a stable design specification. The cost for this tactic is the holding cost for the inventory.

Let us suppose that none of these options is readily viable for FCM. It's too expensive to hire and train a seventh technician; customers will not accept a longer lead time; and the fuel controllers cannot be made to stock as each has some customized content and thus must be made to order. Then another option to reduce overtime might be to increase the current capability of the work force by cross-training. Currently each technician is trained to assemble just one product, and each product can be assembled by just one technician. We term this a *dedicated system*, in which each resource is dedicated to one product and each product can be assembled by only one resource. Suppose we could train and certify the technicians to assemble more types of products. As we will see, this cross-training is a form of risk pooling, as the resource flexibility permits the pooling of the demand variability across multiple products that share the same set of resources; see Chap. 9 in this volume for a discussion of risk pooling.

What is the benefit from having a *flexibility buffer* in the form of cross-trained labor? How is the best way to create this flexibility? We will explore these questions in the remainder of this chapter.

## Total Flexibility

To get started, let us consider an extreme case. Suppose we train each technician to be able to assemble all six products. Again, each unit requires 4 hours of assembly time, so each technician can assemble 10 units in the normal work week. Thus, the work force can assemble 60 units each week without resorting to overtime, regardless of the product mix. This is because the technicians are fully interchangeable, with each of them capable of doing any mix of work — so each can do any 10 units per week. However, overtime will be needed to assemble any demand beyond 60 units.

Under the assumption that demand for each of the six products has outcomes of 8, 10, and 12 units with equal probability, we can find the probability distribution for the total aggregate demand. We report this in Table 3.1 and depict it as a histogram in Fig. 3.2. We also report in Table 3.1 the cumulative probability and overtime required for each demand outcome. For instance, the weekly aggregate demand is 60 units or less with probability 0.5967; also, if demand were 66 units, then overtime is needed to assemble 6 units, requiring a total of 24 h.

From the table we see that overtime is needed for only about 40% of the weeks; in 6 out of 10 weeks, we would not need any overtime. In contrast, in the dedicated system where each technician can assemble only one product type,

**Table 3.1** Probability Distribution for Total Demand

| Demand outcome | Probability | Cumulative probability | Overtime production (units) | Overtime (hours) |
|---|---|---|---|---|
| 48 | 0.0014 | 0.0014 | | |
| 50 | 0.0082 | 0.0096 | | |
| 52 | 0.0288 | 0.0384 | | |
| 54 | 0.0686 | 0.1070 | | |
| 56 | 0.1235 | 0.2305 | | |
| 58 | 0.1728 | 0.4033 | | |
| 60 | 0.1934 | 0.5967 | | |
| 62 | 0.1728 | 0.7695 | 2 | 8 |
| 64 | 0.1235 | 0.8930 | 4 | 16 |
| 66 | 0.0686 | 0.9616 | 6 | 24 |
| 68 | 0.0288 | 0.9904 | 8 | 32 |
| 70 | 0.0082 | 0.9986 | 10 | 40 |
| 72 | 0.0014 | 1.0000 | 12 | 48 |

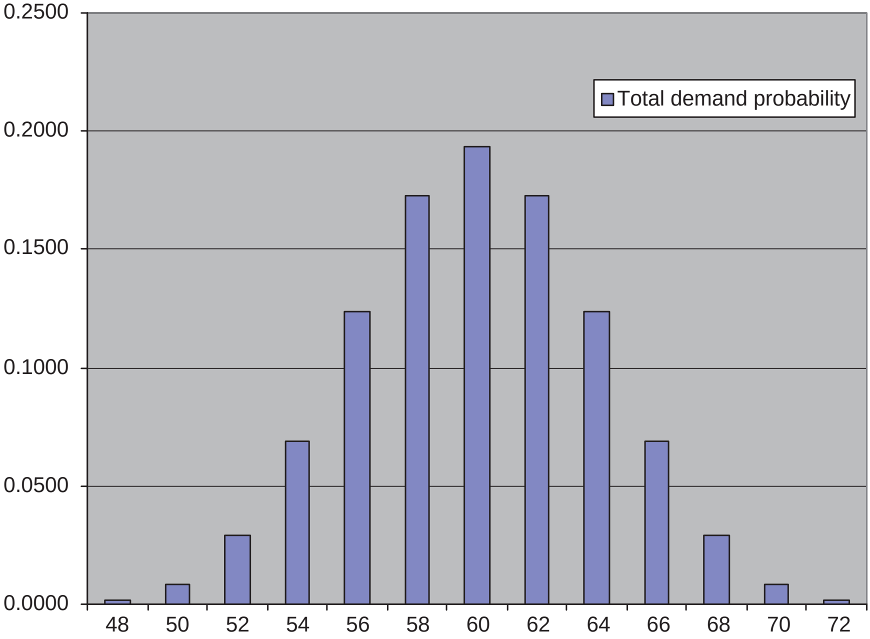


**Fig. 3.2** Total demand probability

then each technician works overtime with probability $\frac{1}{3}$; thus, the probability that there is no overtime needed by any of the technicians would be $\left(\frac{2}{3}\right)^6$ or only 9% of the weeks.

From Table 3.1 we find the average amount of weekly overtime to be 6.32 h. We compare this to the base case in which each technician could assemble only one product; in this case the average overtime is 16 h per week. Hence, for FCM the benefit of complete cross-training of the work force is a reduction in the average overtime from 16 h to 6.32 h per week. This is the absolute best that can be done from increasing the flexibility of the current work force. In the following sections we examine how close we can get to this best case with more selective cross-training.

## Limited Flexibility: Train Adam To Assemble B

As training is quite expensive, a natural question is what the benefits might be from a less ambitious schedule of cross-training. Again, we will explore this by starting with an extreme case. Suppose we train Adam so that he can assemble product B, in addition to product A. All of the other technicians remain dedicated to their products. What is the benefit to FCM from this investment?

With a little thought, one realizes that the only benefit from this investment is when demand for product A is low and demand for product B is high, that is when $d_A = 8$ and $d_B = 12$, where $d_x$ denotes demand for product x. Then Adam can assemble 8 units of A and 2 units of B within his 40-h work week and Bob can assemble the remaining demand for B, namely 10 units, in his normal week. Thus, the cross-training of Adam to produce B would save one shift (8 h) of overtime for Bob whenever $d_A = 8$ and $d_B = 12$.

For all other demand outcomes there is no benefit. When $d_A = 10$ or $12$ , then Adam is of no help to Bob, as he must spend his entire time assembling product A. When $d_A = 8$ and $d_B = 8$ or $10$ , then Adam has idle time but Bob does not need any help.

The probability that $d_A = 8$ and $d_B = 12$ is easily found to be:

$$\Pr\left[d_A = 8 \text{ and } d_B = 12\right] = \Pr\left[d_A = 8\right] \times \Pr\left[d_B = 12\right] = \left(\frac{1}{3}\right) \times \left(\frac{1}{3}\right) = \frac{1}{9}.$$

Thus, relative to the base case of the dedicated system, the benefit from training Adam to assemble B is a reduction in overtime per week, on average, of $\frac{1}{9} \times 8\,h = 0.89\,h$.

For FCM the cost of an hour of overtime is approximately \$80. Thus, the annual benefit from training Adam to assemble B is given by

$$\frac{\$80}{h} \times \frac{0.89h}{week} \times \frac{52 weeks}{year} \approx \frac{\$3700}{year}.$$

FCM needs to compare this to the cost of training; a technician requires between 100 and 150 h of training in order to be certified to assemble a new product. FCM estimates the one-time cost of this to be on the order of \$10,000.

Thus, there would be about a 3-year pay-back from training Adam to assemble B, which is an unacceptable return for FCM. FCM requires a 2-year pay back on this type of investment.

## Limited Flexibility: Create Two-Person Teams

Suppose FCM also trains Bob to assemble product A, so that Adam and Bob can form a two-person team to handle both products A and B. From a similar analysis, we find that there are now two demand scenarios when this cross-training leads to a benefit: $d_A = 8$, $d_B = 12$ and $d_A = 12$, $d_B = 8$. Each scenario occurs with probability $\frac{1}{9}$ and saves 8 h of overtime, relative to the base case in Fig. 3.1. Thus, cross-training both Adam and Bob leads to an average reduction in overtime of $\frac{2}{9}\times 8\,h = 1.78\,h$.

Based on these two examples, one might conclude that cross-training is not a viable option for FCM. Training each technician to be capable to assemble a second product entails a one-time cost of $10,000 per technician. The annual benefit from each cross-trained technician would be a reduction in overtime leading to a savings of $3700 per year. Thus, each investment would be recovered in 3 years, which is not acceptable for FCM. One cross-training plan is shown in Fig. 3.3, in which FCM establishes three two-person teams. This flexibility or cross-training plan results in a reduction in the expected overtime from 16 h per week for the base case to 10.67 h per week.

We previously had found that training each technician to assemble all products reduces the expected overtime from 16 h per week to 6.32 h per week, which is the best we can do. The cross-training plan shown in Fig. 3.3 closes more than half of the gap between the dedicated system (16 h of overtime per week) and the total-flexibility configuration (6.32 h of overtime per week). We can close this gap by

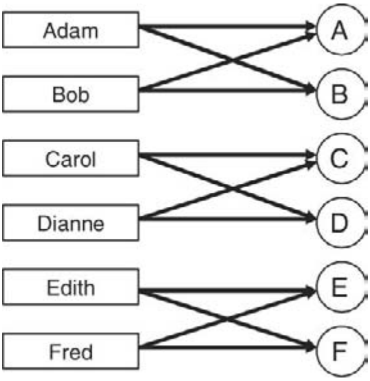


**Fig. 3.3** Two-person teams

adding more flexibility to the configuration in Fig. 3.3, by continuing to cross-train each technician; but one wonders whether there is a better way to proceed.

## Limited Flexibility: Build A Chain

Suppose instead of creating two-person teams we follow a different strategy, as shown in Fig. 3.4. That is, let's train Adam to assemble B, and Bob to assemble C. As before, we see that there are benefits when $d_A = 8, d_B = 12$ and $d_B = 8, d_C = 12$. Each scenario occurs with probability $\frac{1}{9}$ and saves 8 h of overtime, relative to the dedicated system with no cross-training.

But we have another demand scenario to consider here. Suppose demand for product A is low and demand for product C is high, namely $d_A = 8, d_B = 10, d_C = 12$. Then, from inspection, we see that we can accommodate the surplus demand for C without overtime: Adam assembles 8 units of A and 2 units of B; Bob assembles 8 units of B and 2 units of C; and Carol assembles the remaining 10 units of C. Thus, we have been able to use Adam's excess capacity to meet the demand surplus for product C, even though Adam is not capable of assembling product C!

This demand scenario occurs with probability $\left(\frac{1}{3}\right)^3 = \frac{1}{27}$ and saves 8 h of overtime, relative to the base case. Thus, cross-training Adam to product B and Bob to product C leads to an expected reduction in overtime of $\left(\frac{2}{9} + \frac{1}{27}\right) \times 8\ h = 2.07\ h$

Recall that cross-training Adam to product B and Bob to product A led to a reduction of 1.78 h. *Thus, the benefit from cross-training Adam and Bob depends on how it is done.* The plan in Fig. 3.4 is better than a two-person team focused on products A and B. The reason is clear from consideration of the demand scenario $d_A = 8, d_B = 10, d_C = 12$. When demand is low for A but high for C, the configuration in Fig. 3.4 permits Adam's excess capacity to be applied *indirectly* to satisfy the excess demand for C. This is possible because Adam is linked to product C by way of Bob: since Adam can produce B, he can off load work from Bob, who can then use this capacity to help Carol with C.

Furthermore, we observe that the incremental benefit from the second investment in cross-training (training Bob to assemble product C) is *greater* than the benefit for the initial investment (training Adam to assemble B). When FCM trains Adam to assemble B, it obtains a reduction of 0.89 h of overtime per week, for an annual savings of $3,700. Training Bob to assemble C increases the overall benefit to a reduction of 2.07 h per week, for an annual savings in excess of $8,600. Thus, the incremental annual savings from training Bob is more than $4,900! In most contexts, the rule of thumb is that we get decreasing incremental returns from additional investment (for an example, see Chap. 2 in this volume); this is a somewhat surprising counter example to this rule.

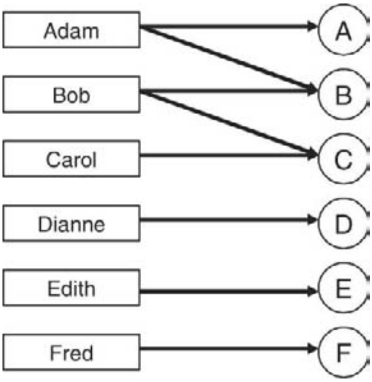


**Fig. 3.4** Chaining Adam, Bob, and Carol

The configuration in Fig. 3.4 is the best way to train two technicians to each do an additional product. In network terminology, this is the best way to add two links or arcs to the dedicated system, shown in Fig. 3.1, where each link signifies the training of one technician to assemble one product. We say that there is a *chain* that connects products A, B, and C to the resources Adam, Bob, and Carol, in that these products and technicians are connected by the links. Within the chain, we can trace a path between any product or technician to any other product or technician using the links[1]. The existence of this chain is what provides the extra benefit, or dividend, relative to the two-person teams in Fig. 3.3: the chain permits Adam's excess capacity to be deployed so as to handle the excess demand at C.

There are many equivalent configurations to that shown in Fig. 3.4, in which we add two links and form a chain. For instance, we could train Adam to assemble C and train Carol to assemble E. This will create a chain from A to Adam, Adam to C, C to Carol, and Carol to E. The benefit is the same as that for Fig. 3.4. When adding two links, the key idea is to coordinate the training so as to create the longest chain.

These ideas extend as we add more links to the network. Consider the cross-training plan shown in Fig. 3.5. To see the benefits, we enumerate all of the relevant demand scenarios in Table 3.2, where blank cells signify that the demand can assume any outcome. From the table we can compute the expected reduction in weekly overtime to be 3.46 h, for an annual savings of nearly $14,400. Thus, the incremental benefit from training Carol to assemble product D is more than $5,700 per year. Again we have an increasing return from the incremental investment in cross-training that is attributable to chaining; we now have a chain connecting product A to product D and technicians Adam to Dianne, which permits the excess capacity of Adam and Bob to be deployed to meet surplus demand for products C and D. Yet, we still do not quite have a viable business case for FCM: the required

[1] In network terminology, a chain is a connected sub-graph of the network.

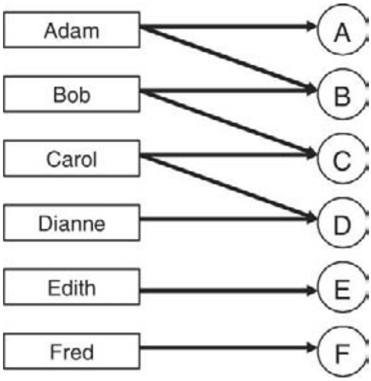


**Fig. 3.5** Chain connecting Adam, Bob, Carol, Dianne with products A, B, C, D

**Table 3.2** Demand Cases and Incremental Benefits from Configuration in Fig. 3.5[2]

| $d_A$ | $d_B$ | $d_C$ | $d_D$ | Probability | Overtime savings (hours) |
|---|---|---|---|---|---|
| 8 | 12 | | | 0.1111 | 8 |
| | 8 | 12 | | 0.1111 | 8 |
| | | 8 | 12 | 0.1111 | 8 |
| 8 | 10 | 12 | | 0.0370 | 8 |
| | 8 | 10 | 12 | 0.0370 | 8 |
| 8 | 10 | 10 | 12 | 0.0123 | 8 |
| 8 | 8 | 12 | 12 | 0.0123 | 8 |

investment is $30,000 to train Adam, Bob, and Carol and the annual savings are less than $15,000, leading to a more than 2-year payback.

## Limited Flexibility: Complete the Chain

We can continue in this fashion and add additional flexibility—train Dianne to assemble E and train Edith to assemble F. We then have a configuration as shown in Fig. 3.6, with a chain that connects all products and all technicians. We find that the incremental savings from training Dianne to assemble product E is $6,300 per year; the incremental savings from then training Edith to assemble F is $6,711 per year. Thus, we find that there continues to be an increasing return from additional cross-training—as long as we keep building a longer chain. Furthermore, from the

[2] The probabilities in the table are just the product of the probabilities for the specified demand outcomes, which are assumed to be independent; for instance, the probability for the demand case in the fifth row is found as:

$$\Pr[\mathrm{d}_B = 8, \mathrm{d}_C = 10, \mathrm{d}_D = 12] = \Pr[\mathrm{d}_B = 8] \times \Pr[d_D = 12] = \left(\frac{1}{3}\right)^3 = \frac{1}{27} = 0.0370.$$

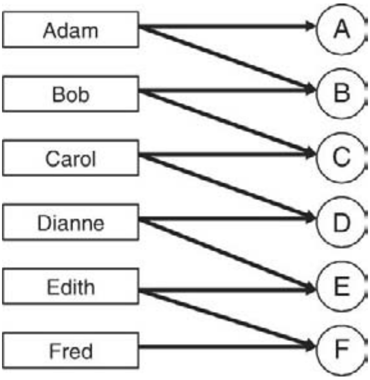


**Fig. 3.6** Chain connecting all products and technicians

standpoint of FCM, the payback for the $10,000 investment in cross-training is now about 18 months for both Dianne and Edith; this is now a quite attractive return on investment for FCM.

However, there is still a big gap between the performance of the chain in Fig. 3.6 (with all products and technicians connected) and that for a total flexibility configuration in which each technician can assemble all products. The expected amount of overtime for the configuration in Fig. 3.6 is 9.42 h per week, whereas it is 6.32 h per week if each technician is fully flexible, capable of assembling any product. Again, we ask whether we can close this gap with limited flexibility, and if so, how?

We note that whereas we have connected all products and technicians in Fig. 3.6, there are some asymmetries in the configuration. There are two technicians that can assemble each product, except for product A that can only be made by Adam; and we have cross-trained all technicians to assemble two products except for Fred. What if we now train Fred to assemble A, as shown in Fig. 3.7?

We find that this last investment in cross-training is extremely beneficial. As shown in Table 3.3 and Fig. 3.8, adding this link, namely training Fred to assemble A, results in an incremental annual savings of over $12,800. Furthermore, we find that the configuration in Fig. 3.7 performs equivalent to the total flexibility configuration. Each configuration averages 6.32 h of overtime per week, for an annual overtime cost of $26,300. Yet, these two systems differ dramatically in terms of their investment in cross-training. On the one hand, the closed chain configuration requires an investment of $60,000, to train each technician to assemble a second product; the savings relative to the base case of a dedicated system is $40,000 per year, leading to payback in 1.5 years. On the other hand, the total flexibility configuration entails training each technician to produce all five other products, say, for a one-time training cost of $50,000 per technician for a total of $300,000 for FCM; the payback for the investment in total flexibility would be 7.5 years.

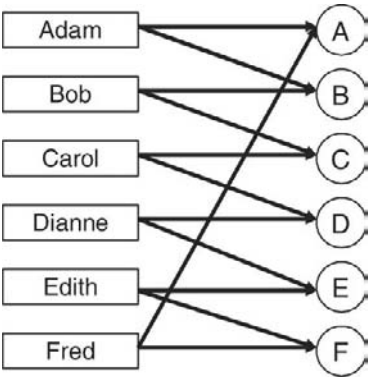


**Fig. 3.7** Closed chain

**Table 3.3** Performance of Chains, Adding One Link at a Time

| Configuration | Expected overtime per week (hours) | Expected cost per year ($ per year) | Incremental reduction from link ($ per year) |
|---|---|---|---|
| Dedicated system | 16.00 | 66,560 | |
| Train Adam on B | 15.11 | 62,862 | 3,698 |
| Train Bob on C | 13.93 | 57,932 | 4,930 |
| Train Carol on D | 12.54 | 52,180 | 5,752 |
| Train Dianne on E | 11.03 | 45,880 | 6,300 |
| Train Edith on F | 9.42 | 39,169 | 6,711 |
| Train Fred on A | 6.32 | 26,295 | 12,874 |
| Total flexibility | 6.32 | 26,295 | |

Why is this last link so valuable? This last link *closes the chain* in that we now have a complete circuit that connects all products and technicians. This circuit or closed chain permits us to move excess capacity from any part of the network to meet surplus demand for any of the products. In particular, if we consider the configuration in Fig. 3.6, even though we have chained all of the products and resources together, there is no way to move excess capacity from the bottom of the network (i.e., Edith or Fred) to meet surplus demand at the top of the network (products A and B); adding the last link from Fred to product A now permits this to happen. Indeed, with the closed chain we now have a remarkable ability to match the available capacity to the variable demand.

Furthermore, in this example any additional cross-training provides no benefit. The closed chain configuration is literally equivalent to the system with total flexibility, for the given demand outcomes. Having more cross-training does not lead to any reductions in the overtime requirements. Thus, this is the best we can do.

As another way to see the value from chaining, let’s compare the plan in Fig. 3.3 (two-person teams) with the plan in Fig. 3.7 (closed chain). In both instances we train each technician to be able to assemble a second product. In both instances, each product can be assembled by two technicians. These two systems require the same investment in cross-training, $60,000, and would seem to be very similar in structure.

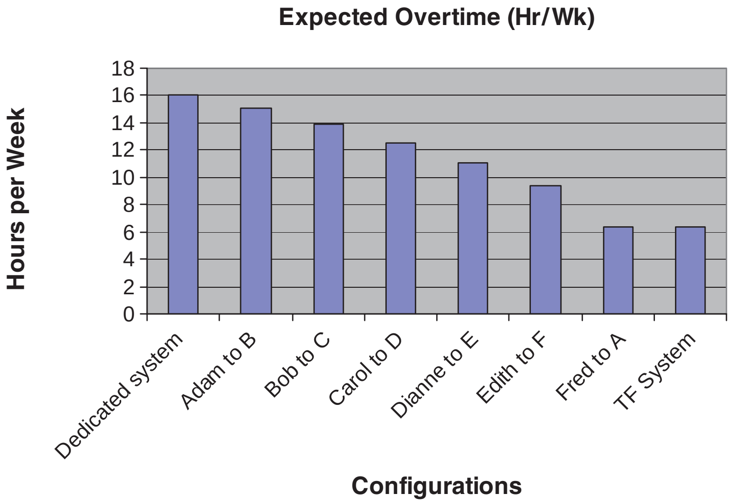


**Fig. 3.8** Results from adding flexibility to base case

But the return on this investment is much different, as these two systems perform much differently. In comparison with the dedicated system (Fig. 3.1), the plan with two-person teams reduces the weekly overtime, on average, from 16 h to 10.67 h. In contrast, the plan with the closed chain reduces the weekly overtime, on average, from 16 h to 6.32 h, yielding almost twice the benefit from the two-person teams.

## General Insights On Flexibility

We have developed and used the FCM example as a way to examine flexibility in an operating setting. In this case, we have a relatively simple production system with uncertainty in the demand requirements and with production inflexibility due to constrained resources. We have an opportunity to increase the flexibility of the system by cross-training; however, there is a trade-off between the cost of cross-training and the possible benefits from reduced overtime. From this example we observe the following:

- The benefits from increased flexibility depend on how it is deployed; in particular, we get the greatest benefits by building longer and longer chains that connect more and more products and resources together.
- As we build a chain, we get increasing returns from our investments in flexibility; that is, the incremental benefit from an investment in flexibility grows as we increase the size of the chain.
- Once we have created a chain that connects all of the products and resources together, we get the largest incremental benefit by closing the chain; that is, we get the largest return by adding a link that creates a circuit connecting all of the products and resources.

- In our example the performance of the closed chain configuration is equivalent to that for a system with total flexibility, that is, a system in which each resource can produce any of the products. Thus, there is no benefit from adding any flexibility beyond the closed chain.

Although we have developed these observations for a quite stylized setting, we have found them to be quite robust. Indeed, we contend that (1) limited flexibility, when deployed in the right way, yields most of the benefits of total flexibility, and (2) limited flexibility provides the greatest benefits when configured to chain products and resources together to the greatest extent possible. In particular, as we increase the demand variability for the products, we do find that the performance of a totally flexible system can be better than that for a system with a closed chain configuration. However, for realistic ranges of the demand variability, this performance gap, albeit not zero as in the case of the example, is quite small. As a consequence, we observe that there is usually a very poor return from adding any flexibility beyond the closed chain.

The FCM example has a great amount of symmetry—same number of products and resources, same production rates for each resource, same demand characteristics for each product. This is not the case for more realistic scenarios, where each product has distinct demand characteristics, the production rates vary across the resources and the number of products differs from the number of resources. Nevertheless, we have found that the principles cited above are most helpful in providing guidelines for exploring flexibility options. We still want to build as long a chain as possible and to create a circuit that connects the primary products and resources together. In addition we try to equalize the load or expected demand that is assigned to each resource, and try to connect each product to its share of capacity. With these additional considerations, limited flexibility deployed to create chains provides most of the benefits of total flexibility.

The principles that we uncovered with the FCM example apply to many other contexts. The general conditions are that we have multiple resources and multiple tasks. A resource is required to perform each task. However, some form of investment, qualification or training is required in order for a specific resource to have the capability to perform a specific task. As a consequence, each resource has limited capability in that it is qualified to perform only a subset of all of the tasks. Furthermore there is some uncertainty in demand requirements, namely how much work is required for each type of task. Key issues for such an operation are to decide what capability each resource should have and to predict the performance of a given configuration of resources.

## Applications

The FCM example is one context. The multiple resources are highly skilled technicians. The tasks are the assembly of a complex product, e.g., fuel controllers. There is uncertainty in the demand requirements for these products, and there is a nontrivial expense to train a technician to be certified to assemble each of the products. Other examples include

- Maintenance personnel and repair tasks. Equipment in a plant fails at random intervals. The repair task is specific to the type of equipment and requires a qualified maintenance engineer. The flexibility questions are to decide the training for each maintenance engineer, and then how to dynamically assign to repair tasks.
- Service representatives and/or technicians in a call center. A call center receives a mix of types of calls: questions about a product warranty, questions about how to use a product or how to install software, questions about how to return a product for repair, requests to place an order or make a reservation, questions about order status or to cancel/change an order, etc. Service agents handle these calls but require specialized skills and/or access to on-line information, depending upon the product and the nature of the call. Most call center operating systems will automatically screen calls to determine the type of call so as to direct it to an appropriate service agent. The flexibility questions are to decide how many service agents and their training, and how to route calls as they arrive to the system.
- An assembly cell. An assembly cell is set up to assemble a product or product family. The assembly of a product is broken down into a sequence of specific tasks. The assembly cell consists of a sequence of workstations, at each of which a subset of the tasks are to be performed. A trained technician is required to perform the tasks at each workstation, but often a cell will have fewer technicians than work stations. In this case, technicians will float between workstations. The flexibility questions are to decide how many technicians and what amount of cross-training is needed to assure an efficient operation.
- Semiconductor fabrication. The production of a semiconductor device entails several hundred process steps that are necessary to build twenty to thirty layers of circuitry. For many of the process steps, a device returns to the same set of process tools for each layer. Each process tool is capable of performing the process task for any layer, but usually requires a setup and/or specialized tooling to switch from one layer to another layer. The flexibility questions are to decide which tools to dedicate to which layers, which tools to flex between multiple layers, and then how to schedule the work flow through the set of process tools.
- Assembly plants. In the automobile industry, the assembly plants for final vehicle assembly can produce multiple vehicle models or name plates. But such flexibility requires extensive planning and investment in tooling and training. The flexibility questions are to decide for a set of assembly plants and a set of vehicle models, which models are to be produced in which plants.

## Historical Background

The key reference for this chapter is *Jordan and Graves (1995)* who examine the benefits of process flexibility in the context of a set of plants producing a set of products. They introduce the concept of chaining and establish that (a) limited flexibility will yield nearly all of the benefits of total flexibility, if configured in the right way, and that (b) limited flexibility provides the greatest benefits when configured to chain as many plants and products together as possible.

There have been a number of important subsequent developments. We mention a few here.

Hopp et al. (2004) examine cross-training in a serial production or assembly line and establish the value of chaining; they find that a cross-training strategy that attempts to chain together workers and workstations performs best.

Gurumurthi and Benjaafar (2004) consider a queuing system with multiple customer classes and heterogeneous servers. Each server can have the capability to process more than one customer class. They develop a framework and computational algorithm for evaluating the performance of different flexibility configurations and control policies. They find that limited flexibility, in the form of chaining, works extremely well in many settings.

Graves and Tomlin (2003) extend the model from Jordan and Graves to a multistage supply chain in which each product requires processing at each stage of the supply chain. They show how the single-stage guidelines from Jordan and Graves adapt and apply to a multistage setting.

Jordan et al. (2004) and Inman et al. (2004) examine two contexts in which cross-training can arise in the automobile industry. Jordan et al. (2004) consider the value of cross-training a workforce to perform maintenance and repair tasks. Inman et al. (2004) consider how best to do cross-training for an assembly line to counteract the impact from absenteeism. In both case, they find that a chaining strategy is effective.

**Building Intuition** The performance of a service or manufacturing system with multiple products being served by multiple resources depends upon the flexibility of the resources. A resource is flexible if it is capable of serving several products, but making a resource flexible can be quite expensive. Hence there is always a tradeoff between the costs and benefits from increasing the flexibility in these systems.

We show by example that limited flexibility, when deployed in the right way, yields most of the benefits of total flexibility. Furthermore, limited flexibility provides the greatest benefits when configured to chain products and resources together to the greatest extent possible.

For realistic settings, these principles provide helpful guidelines for exploring flexibility options. We still want to build as long a chain as possible and to create a circuit that connects the primary products and resources together. In addition we try to equalize the load or expected demand that is assigned to each resource, and try to connect each product to its share of capacity. With these considerations, limited flexibility deployed to create chains provides most of the benefits of total flexibility.

Iravani et al. (2005) consider a general setting with multiple parallel resources and multiple parallel tasks. They define a configuration by a specification of the capabilities of each resource, i.e., which tasks each resource can perform. They develop and examine various measures for predicting the flexibility of a configuration, and provide additional evidence for the value of chaining.

Wallace and Whitt (2005) consider cross-training of agents in a call center with several types of calls. They discover that when each agent has only two skills, the system performance can be comparable to that when each agent has every skill, provided that the cross-training is done the right way. Again, the concept of chaining together as many agents and skills as possible is the key insight.

**Acknowledgements** This research has been supported in part by the MIT Leaders for Manufacturing Program, a partnership between MIT and major manufacturing firms; and by the Singapore-MIT Alliance, an engineering education and research collaboration among the National University of Singapore, Nanyang Technological University, and MIT.

## Selected Bibliography

Graves, S. C. and B. T. Tomlin (2003), “Process Flexibility in Supply Chains,” *Management Science*, 49 (7), 907–919.

Gurumurthi, S. and S. Benjaafar (2004), “Modeling and Analysis of Flexible Queueing Systems,” *Naval Research Logistics*, 51 (5), 755–782.

Hopp, W. J., E. Tekin, and M. P. Van Oyen (2004) “Benefits of Skill Chaining in Serial Production Lines with Cross-Trained Workers,” *Management Science*, 50 (1), 83–98.

Inman, R. R., W. C. Jordan and D. E. Blumenfeld (2004) “Chained Cross-Training of Assembly Line Workers,” *International Journal of Production Research*, 42 (10), 1899–1910.

Iravani, S. M., M. P. Van Oyen and K. T. Sims (2005) “Structural Flexibility: A New Perspective on the Design of Manufacturing and Service Operations,” *Management Science*, 51 (2), 151–166.

Jordan, W. C. and S. C. Graves (1995), “Principles on the Benefits of Manufacturing Process Flexibility,” *Management Science*, 41 (4), 577–594.

Jordan, W. C., R. R. Inman, and D. E. Blumenfeld (2004), “Chained Cross-Training of Workers for Robust Performance,” *IIE Transactions*, 36 (10), 953–967.

Wallace, R. B. and W. Whitt (2005) “A Staffing Algorithm for Call Centers with Skill-Based Routing,” *Manufacturing & Service Operations Management*, 7 (4), 276–294.

# Chapter 4
# Single Server Queueing Models

**Wallace J. Hopp**
**University of Michigan**

*Queues or waiting lines form in systems when service times and arrival rates are variable. Simple queueing models provide insight into how variability subtly causes congestion. Understanding this is vital to the design and management of a wide range of production and service systems.*

## Introduction

Macrohard is a small startup software company that sells a limited array of products. Like many technology product firms, Macrohard provides technical support to its customers via a toll free number that is available 24 h a day, 7 days a week. However, because of its small size, Macrohard can only justify having a single technician staffing the call center at any given time. This technical support center is vital to Macrohard's business strategy, since it has a significant impact on customers' experience with their products. To ensure that this experience is positive, technicians must provide accurate information and helpful support. From Macrohard's perspective, this is a matter of ensuring that the technician has the right skill and training. But for customers to be happy, service also has to be responsive. A long wait on hold might turn an otherwise satisfied customer into an angry noncustomer.

The Macrohard manager responsible for technical support is therefore interested in a variety of questions concerning customer waiting, including:

1. What arrival rate can a single technician reasonably handle without causing excessive waiting?
2. How likely is it that a customer will have to wait on hold?
3. How long can we expect customers to have to wait on hold?
4. What factors affect the likelihood and duration of customer waiting?
5. What options exist for reducing customer waiting time?

To say anything about these questions we will clearly need some data on the technical support system. One key piece of information is how fast the technician

D. Chhajed and T.J. Lowe (eds.) *Building Intuition*: *Insights From Basic Operations Management Models and Principles*.
doi:10.1007/978-0-387-73699-0, © Springer Science + Business Media, LLC 2008

can process calls, subject to constraints on accuracy, completeness, and politeness. Suppose for the purposes of discussion that Macrohard has timed calls in the past and has found that the average time to handle a call is 15 min.

Now let us appeal to our intuition and see if we can address any of the above questions.

*What arrival rate can a single technician reasonably handle without causing excessive waiting?* At first blush, question (1) appears to be the simplest. Since it takes 15 min per customer, the technician should be able to provide service to four customers per hour. If more than four customers per hour call for support, the technician will be overloaded and people will have to be put on hold. If less than four calls per hour come in, the technician should have idle time. Hence, the technician should be able to service all customers without waiting as long as the arrival rate is less than four per hour. Right?

Wrong! The above logic only makes sense if service times are exactly 15 min for every customer and calls come in perfectly evenly spaced. But we have not assumed that this is the case, and such behavior would be very unlikely in a call center. We would expect some callers to have easy questions, and hence require short service times, while other callers have difficult questions, which require lengthy service times to handle. Furthermore, since customers make independent decisions on when to call in, we would hardly expect uniform spacing of calls.

For example, assume that the average service time is 15 min and the arrival rate is three calls per hour (one call every 20 min), but that there is some variation in both service times and interarrival times. Now suppose customer A calls in, finds the technician idle, and has a 17-min conversation. Further suppose that customer B calls in 15 min after customer A. This entirely plausible sequence of events results in customer B waiting on hold for 2 min. Hence, even though the technician is fast enough to keep up with calls on average, fluctuations in either the arrival rate or service rate can lead to backups and hence waiting.

The realization that waiting can occur even when the technician has more than enough capacity to keep up with demand leads us to the other four questions in the above list. How likely is a delay? How long will it be? And so on. Unfortunately, intuition abandons us entirely with respect to these questions. Simply knowing that a burst of calls will lead to customers having to wait on hold does not give us a clue of how long waiting times will be on average.

This lack of intuition into the causes and dynamics of waiting is not limited to call center management. Waiting is everywhere in modern life. We wait at Starbucks for coffee to start the day. Then we wait at the toll booth on the drive to work. At the office, we wait to use the copier, wait for a technician to resolve a computer problem and wait for our meeting with the boss. On the way home, we wait at stop lights and wait at the checkout to buy groceries. In the evening we wait for a file to download from the web. If (heaven forbid) our schedule includes a visit to a medical care facility, government office, or an airport, then waiting may well be the dominant activity of the day.

These and thousands of other everyday occurrences are examples of **queueing**, which is the technical (or at least English) term for waiting. Because it is so prevalent, understanding queueing is part of understanding life. But, there are also plenty of not-so-mundane examples of queueing, including ambulance service, organ transplants, security checkpoints, and many more. So an understanding of queueing is also enormously useful in designing and managing all kinds of production and service systems.

**Building Intuition** In any system where entities (customers, jobs, orders, etc.) are processed sequentially, waiting lines form when service rates do not keep up with arrival rates. This will occur occasionally if service times and interarrival times are random and unsynchronized.

The busier the server, the more frequent such backups will be. Hence, variability and utilization combine to cause congestion in sequential processing systems. This also implies that the server cannot always be busy because 100% utilization in a system with variability would have an infinite queue.

Finally, the interaction between variability and utilization in causing waiting lines means that variability reduction can serve as a substitute for capacity increases. Smoothing arrivals and/or service times can enable a system to operate at higher utilization while keeping queue times to acceptable levels.

These insights are central to the design of production and service systems that are both efficient and responsive.

Unfortunately, queueing behavior is subtle. Unlike many other phenomena, waiting phenomena cannot be understood in terms of average quantities. For example, the time it takes to drive between two points depends only on the distance and the *average* driving speed. We do not need to know whether the trip was made at constant speed, or whether the driver sped up and slowed down, or even how often the driver stopped. Average speed is all that matters. Similarly, the amount of nondefective product produced by a factory during the month depends only on the amount produced and the yield rate (i.e., fraction of product that passes a quality test). We do not need to know whether some days had higher yields than others. Average yield rate is all that matters.

However, queueing behavior depends on more than average quantities. As we argued above, merely knowing that the service rate is four calls per hour and the call rate is three calls per hour in the Macrohard system is not sufficient for us to predict the amount of waiting in the system. The reason is that **variability** of arrivals and service times also contribute to waiting. Hence, in order to understand queueing behavior we must characterize variability and describe its role in causing waiting to occur.

Questions about queueing behavior, in the Macrohard system and elsewhere, are clearly more difficult than the question of how long it will take to drive a given distance. So answering them will require a more sophisticated model than a simple ratio of distance over time. In this chapter we examine models of the simplest queueing systems, namely those with a single server, in order to develop intuition into the causes and nature of waiting. We will also develop the analytic tools needed to answer the previous list of questions concerning the Macrohard system.

## Queueing Systems

In order to model queueing systems, we first need to be a bit more precise about what constitutes a queueing system. The three basic elements common to all queueing systems are:

**Arrival Process**: Any queuing system must work on something — customers, parts, patients, orders, etc. We generically term these **entities**. Before entities can be processed or subjected to waiting, they must first enter the system. Depending on the environment, entities can arrive smoothly or in an unpredictable fashion. They can arrive one at a time or in clumps (e.g., bus loads or batches). They can arrive independently or according to some kind of correlation (e.g., a network server that processes jobs from only five remote computers will have correlated arrivals, since if three of the computers are already waiting for a response from the server, we know that only the two remaining computers can make requests). In this chapter we assume that entities arrive one at a time and that times between arrivals are uncorrelated with each other and with the service times. We also assume that all arrivals enter the system (i.e., do not balk) and remain in the system until serviced (i.e., do not reneg).

A special arrival process, which is highly useful for modeling purposes, is the **Markov** arrival process. In this process, entities arrive one at a time and the times between arrivals are **exponential** random variables. This type of arrival process is *memoryless*, which means that the likelihood of an arrival within the next $s$ minutes is the same no matter how long it has been since the last arrival. For example, suppose an arrival has just occurred (at time $t = 0$) and that the probability of another arrival within the next 5 min is 0.25. Furthermore, suppose that no arrivals occur for 10 min. If the arrival process is Markov, then the probability of an arrival occurring within the next 5 min (i.e., by time $t = 15$) is still 0.25. This may seem counterintuitive if one thinks of smooth scheduled arrivals. But if one thinks of unpredictable public transit systems (e.g., buses), it may not be to much of a stretch to believe that the likelihood of an arrival does not increase as one waits. Moreover, there are theoretical results showing that if a large population of customers makes independent decisions of when to seek service, the resulting arrival process will be Markov. Examples where this occurs are phone calls arriving at an exchange, customers arriving at a fast food restaurant, hits on a web site, and many others. We will see below that a Markov arrival process leads to tractable models of queueing systems.

**Service Process**: Once entities have entered the system they must be served. The physical meaning of “service” depends on the system. Customers may go through the checkout process. Parts may go through machining. Patients may go through medical treatment. Orders may be filled. And so on. From a modeling standpoint, the operational characteristics of service matter more than the physical characteristics. Specifically, we care about whether service times are long or short, and whether they are regular or highly variable. We care about whether entities are serviced by a single server or by multiple servers working in parallel. We care about whether entities are processed in first-come-first-serve (FCFS) order or according

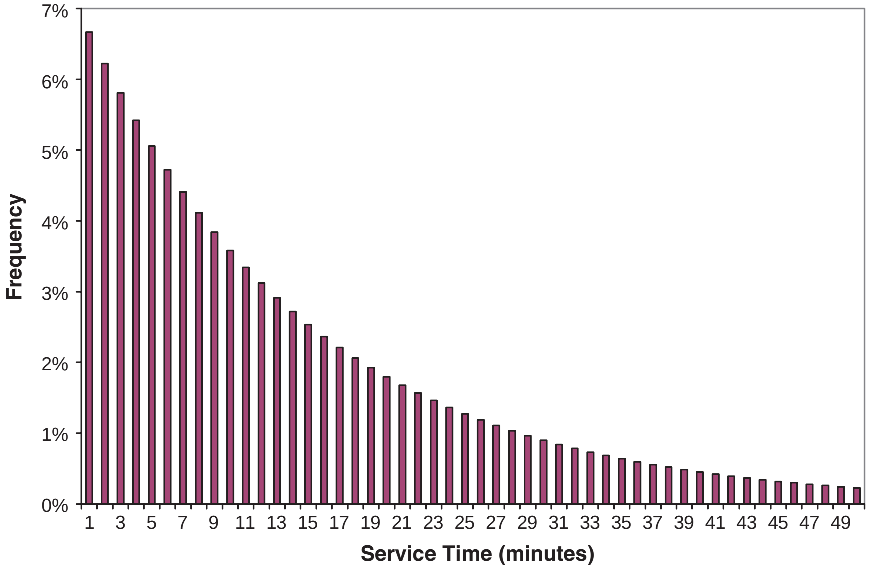


**Fig. 4.1** Exponential distribution with mean of 15 min

to some kind of priority rule. These and the many other operational variations possible in service processes make queueing a very rich subject for modeling research. In this chapter, we assume that entities are served one at a time and that service times are uncorrelated with each other and with the times of arrivals (e.g., the server does not speed up when many customers are waiting).

A special service process is the **Markov** service process, in which entities are processed one at a time in FCFS order and service times are independent and **exponential**. As with the case of Markov arrivals, a Markov service process is memoryless, which means that the expected time until an entity is finished remains constant regardless of how long it has been in service. For example, in the Marcrohard example, a Markov service process would imply that the additional time required to resolve a caller's problem is 15 min, no matter how long the technician has already spent talking to the customer. While this may seem unlikely, it does occur when the distribution of service times looks like the case shown in Fig. 4.1. This depicts a case where the average service time is 15 min, but many customers require calls much shorter than 15 min (e.g., to be reminded of a password or basic procedures) while a few customers require significantly more than 15 min (e.g., to perform complex diagnostics or problem resolution). Simply knowing how long a customer has been in service does not tell us how much more time will be required.

**Queue**: The third required component of a queueing system is a queue, in which entities wait for service. The simplest case is an unlimited queue, which can accommodate any number of customers. But many systems (e.g., phone exchanges, web servers, and call centers), have limits on the number of entities that can be in queue at any given time. Arrivals that come when the queue is full are rejected (e.g., customers

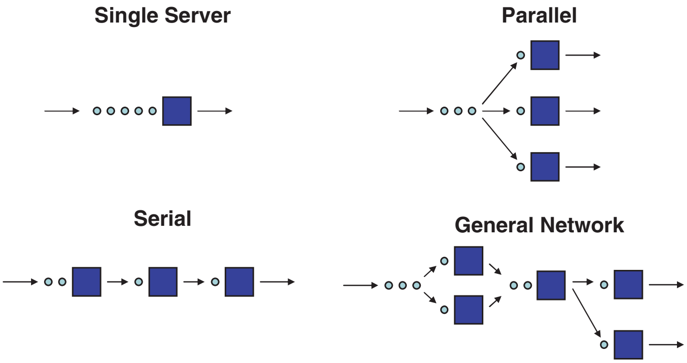


**Fig. 4.2** Queueing network structures

get a busy signal when trying to dial into a call center). Even if the system does not have a strict limit on the queue size, customers may balk at joining the queue when it is too long (e.g., cars pass up a drive through restaurant if there are too many cars already waiting). Entities may also exit the system due to impatience (e.g., customers kept waiting too long at a bank decide to leave without service) or perishability (e.g., samples waiting for testing at a lab spoil after some time period).

In addition to variations in the arrival and service processes and the queueing protocol, queueing systems can differ according to their basic flow architecture. Figure 4.2 illustrates some prototypical systems. The **single server system**, which is the focus of this chapter, is the simplest case and represents many real world settings such as the checkout lane in a drugstore with a single cashier or the CPU of a computer that processes jobs from multiple applications. If we add cashiers or CPUs to these systems, they become **parallel systems** (i.e., a single waiting line serviced by multiple identical servers). Many call centers, bank service counters, and toll booths are also parallel queueing systems.

Waiting also occurs in systems with multiple stages. For example, a production line that fabricates products in successive steps, where each step is performed by a single machine workstation, is an example of a **serial system**. More elaborate production systems, with parallel machine workstations and multiple routings form **general queueing networks**. Modeling and analysis of such networks can become very complicated. But the basic behavior that underlies all of these queueing systems is present in the single server case. Hence, understanding the simple single server queue gives us powerful intuition into a vast range of systems that involve waiting.

## M/M/1 Queue

We now turn our attention to modeling the single server queueing system, like that represented by the Macrohard technical support center.

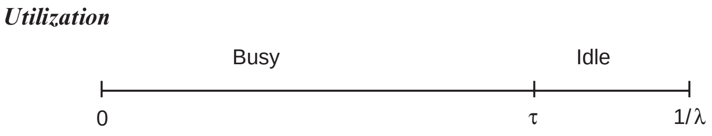


**Fig. 4.3** Fraction of time server is busy

To develop a mathematical model, we let λ (lambda) represent the **arrival rate** (calls per hour) and τ (tau) represent the **average service time** (in hours). In the Macrohard example, τ = 0.25 h (15 min). For convenience, we also denote the **service rate** by μ (mu), which is defined as μ = 1/τ. In the Macrohard call center, μ = 1/15 = 0.067 calls per minute = 4 calls per hour. This represents the rate at which the system can process entities if it is never idle and is therefore called the **capacity** of the system.

## *Utilization*

This above information is sufficient for computing the average fraction of time the server is busy in a single-server queueing system. We do this by noting that the average time between arrivals (calls) is 1/λ. The average amount of time the server is busy during this interval is τ (see Fig. 4.3). Hence, the fraction of time the server is busy, called the **utilization** of the server and denoted by ρ (rho), is given by

$$\rho = \frac{\tau}{1/\lambda} = \lambda\tau = \frac{\lambda}{\mu}. \tag{1}$$

This only makes sense, however, if $\tau < 1/\lambda$, or equivalently $\lambda\tau = \rho < 1$. If this is not the case, then utilization is above 100% and the queueing system will not be stable (i.e., customers will pile up so that the queue grows without bound over time) unless customers balk at joining a queue if it is too long or leave the queue without service if they wait too long.

For example, in the Macrohard service center calls must come in at a rate below four per hour, because this is the maximum rate at which the technician can work. If calls are received at an average rate of, say, three per hour, then server utilization will be ρ = λτ = (3 per hour)(0.25 h) = 75%. This means that the technician will be idle 25% of the time. One might think that customer waiting would be almost nonexistent, since the technician has so much extra time. But, as we will see below, this is not necessarily the case.

## *Birth-Death Process Model*

The average arrival and service rates are not enough to allow us to compute the average time an entity waits for service in a queueing system. To do this, we must introduce variability into arrival and service processes. The most straightforward

***Birth-Death Process Model***

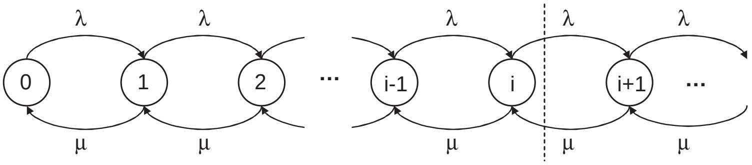


**Fig. 4.4** Birth–death process model of M/M/1 queue

queueing model that incorporates variability is the single server system with Markov (memoryless) arrival and service processes. We call this the M/M/1 queue, where the first M indicates that the arrival process is Markov, the second M indicates that the service process is Markov and the 1 indicates that there is only a single server. Waiting space is assumed infinite, so there is no limit on the length of the queue. We will examine some other single server queues later in this chapter.

The assumption that both the arrival and service processes are memoryless means that the only piece of information we need to predict the future evolution of the system is the number of entities in the queue. For example, in the Macrohard support center if there is currently one customer being served and one waiting on hold, then we know that the expected time until the customer completes service is 15 min and the expected time until the next customer calls for service is 20 min. Because both interarrival and service times are exponentially distributed, and hence memoryless, we do not need to know how long the current customer has been in service or how long it has been since a call came into the system.

This fact allows us to define the number of customers in the queue as the system **state**. Moreover, since arrivals come into the system one at a time and customers are also serviced one at a time, the state of the system either increases (when a customer arrives) or decreases (when a service is completed) in increments of one. Systems with this property are called **birth–death processes**, because we can view increases in the state as “births” and decreases as “deaths.” (Note that birth–death processes do not allow “twins,” since increases and decreases are assumed to occur in increments of one.) Figure 4.4 shows the progression of state in the birth–death process model of the M/M/1 queue.

We can use the birth–death representation of the M/M/1 queue to compute the long-run average probabilities of finding the system in any given state. We do this by noting that in steady state, the average number of transitions from state $i$ to state $i + 1$ must equal the average number of transitions from state $i + 1$ to state $i$. Graphically, the number of times the birth–death process in Fig. 4.4 crosses the vertical line going left to right must be the same as the number of times it crosses the vertical line going right to left. Intuitively, the number of births must equal the number of deaths over the long term.

Letting $p_i$ represent the average fraction of time there are $i$ customers in the system, the average rate at which the process crosses the vertical line from left to right is $p_i\lambda$, while the average rate at which it crosses it from right to left is $p_{i+1}\mu$. Hence,

$$p_{\mathrm{i}}\lambda = p_{\mathrm{i+1}}\mu, \text{ and}$$

$$p_{i+1} = \frac{\lambda}{\mu} p_i = \rho\, p_i.$$

This relationship holds for $i = 0,1,2,\ldots$, so we can write

$$p_1 = \rho p_0,$$
$$p_2 = \rho p_1 = \rho^2 p_0,$$
$$p_3 = \rho p_2 = \rho^3 p_0,$$
$$\ldots$$
$$p_{\mathrm{n}} = \rho p_{n\text{-}1} = \rho^n p_0,$$
$$\ldots$$

If we knew $p_0$ (i.e., the probability of an empty system) we would be able to compute all of the other long-term probabilities. To compute this, we note that since the system must be in some state, the long-term probabilities must sum to one. This implies the following

$$1 = \sum_{n=0}^{\infty} p_n = \sum_{n=0}^{\infty} \rho^n p_0 = \frac{p_0}{1-\rho}. \tag{2}$$

Note that the infinite sum in Eq. (2) converges only if $\rho = \lambda/\mu < 1$. That is, the arrival rate ($\lambda$) must be less than the service rate ($\mu$) in order for the queue to be stable and have long term probabilities. If $\rho \geq 1$, then the queue length will grow without bound and hence the system will not be stable.

Equation (2) implies that the probability of finding the system empty is

$$p_0 = 1 - \rho, \tag{3}$$

and hence the long-term probability of being in state $n$ is given by

$$p_{\mathrm{n}} = (1-\rho)\rho^{\mathrm{n}}, \quad n = 0,1,2,\ \ldots,\ k. \tag{4}$$

This is known as the **geometric distribution**, whose mean is given by

$$L^{M/M/1} = \sum_{n=0}^{\infty} n p_n = \sum_{n=0}^{\infty} n(1-\rho)\rho^n = \frac{\rho}{1-\rho}. \tag{5}$$

So, $L^{M/M/1}$ represents the average number of customers in the system (i.e., in service or waiting in queue). Again, this is only well-defined if ρ < 1.

Finally, we note that the expected number of customers in service is given by the probability that there are one or more customers in the system, which is equal to $1-p_0$ (i.e., one minus the probability that the system is empty.) Hence, the **average number in queue** can be computed as the expected number of customers in the system minus the expected number in service, which is

$$L_q^{M/M/1} = L^{M/M/1} - (1 - p_0) = \frac{\rho}{1-\rho} - \rho = \frac{\rho^2}{1-\rho}. \tag{6}$$

We can apply Little's Law (see Chap. 5) to compute the expected waiting time in the system and in the queue as follows:

$$W^{M/M/1} = \frac{L^{M/M/1}}{\lambda} = \frac{\rho}{\lambda(1-\rho)} = \frac{\tau}{(1-\rho)}, \text{ and} \tag{7}$$

$$W_q^{M/M/1} = \frac{L_q^{M/M/1}}{\lambda} = \frac{\rho^2}{\lambda(1-\rho)} = \left(\frac{\rho}{1-\rho}\right)\tau. \tag{8}$$

## *Example*

The above results give us the tools to answer many of the questions we raised previously concerning the performance of the Macrohard technical support system, at least for the situation where both arrivals and services are Markov. For purposes of illustration, we assume that the mean time for the technician to handle a call is 15 min (τ = 0.25 h) and, initially, that the call rate is three calls per hour (λ = 3 per hour). Hence, system utilization is ρ = λτ = 3(0.25) = 0.75. Now let us return to the questions.

*How likely is it that a customer will have to wait on hold?* From Eq. (3) we know that the probability of an empty system is $p_0 = 1-\rho$. Since the system must either be empty or busy at all times, the probability of a busy system is $1-p_0 = \rho$. Thus, the likelihood of a customer finding the system busy is equal to the fraction of time the technician is busy. Since ρ = 0.75, customers have a 75% chance of being put on hold in this system.

*How long can we expect customers to have to wait on hold?* Since the average service time is τ = 15 min, we can use Eq. (8) to compute the average waiting time (i.e., time spent on hold) to be:

$$W_q^{M/M/1} = \left(\frac{\rho}{1-\rho}\right)\tau = \left(\frac{0.75}{1-0.75}\right)0.25 = 0.75 \text{ h} = 45 \text{ min.}$$

Even though the technician is idle 25% of the time, the average waiting time for a customer on hold is 45 min! The reason is that variability in the customer arrivals (and technician service times) results in customers frequently arriving when the technician is busy and therefore having to wait for service.

*What arrival rate can a single technician reasonably handle without causing excessive waiting?* We have already argued that the fact that the capacity of the technician is $\mu = 4$ calls per hour does not mean that the call center can actually provide service to four customers per hour. The reason is that if we had an arrival rate of $\lambda = 4$ customers per hour, the utilization would be $\rho = 1$ and hence the average queue length would be $L = \infty$, which is clearly impossible.

Since we are assuming Markov arrivals and services, this queueing system has variability and hence can only attain its capacity with an infinite queue, and thus infinite waiting time. Therefore, the volume of customers the system can handle in reality is dictated by either the queue space (e.g., number of phone lines available for keeping customers on hold) or the patience of customers (e.g., the maximum time they will remain on hold before hanging up).

Suppose that Macrohard makes a strategic decision that customers should be kept on hold for no more than 5 min on average. This will be the case if the following is true:

$$W_q^{M/M/1} = \frac{\rho}{1-\rho}\tau = \frac{\rho}{1-\rho}15 < 5, \text{ or}$$

$$\frac{\rho}{1-\rho} < \frac{1}{3}, \text{ or}$$

$$\rho < \frac{1}{4}.$$

That is, the customer wait in queue (on hold) will be less than 5 min only if the technician is utilized less than 25%. This places the following constraint on the arrival rate:

$$\rho = \lambda\tau = \lambda(0.25\text{ h}) < 0.25, \text{ or}$$

$$\lambda < 1/\text{h}.$$

Not surprisingly, if we want utilization to be below 25%, the arrival rate must be less than one quarter of the service rate. The implication is that the (not unreasonable) constraint that customers should not be required to wait on hold for more than 5 min on average means that the system can only handle one call per hour, which is far below its theoretical capacity.

Of course, if Macrohard management was willing to let customers wait on hold for more time, the service center could handle a higher call rate. In general, if we want the average waiting time to be no more than *t* minutes, then we can compute the maximum allowable arrival rate ($\lambda$) as follows:

$$W_q^{M/M/1} = \frac{\rho}{1-\rho}\tau = \frac{\rho}{1-\rho}0.25 < t, \text{ or}$$

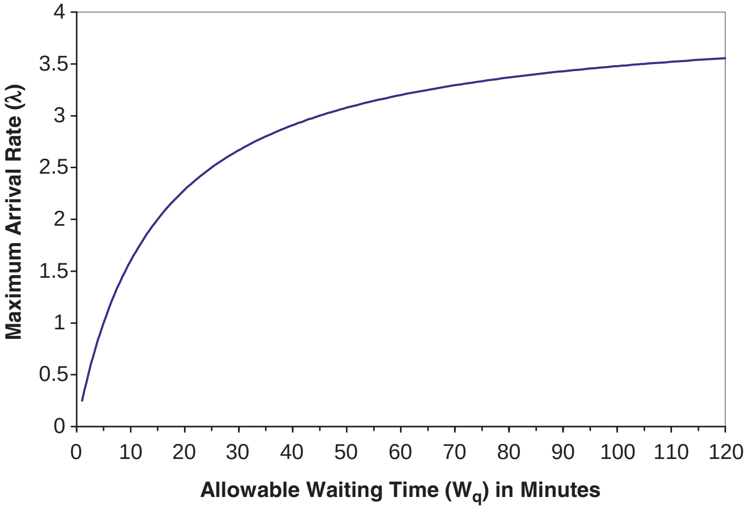


**Fig. 4.5** Maximum arrival rate as a function of allowable waiting time

$$0.25\rho < \mathrm{t} – \mathrm{t}\rho, \text{ or}$$

$$\rho = \lambda(0.25) < \frac{t}{0.25+t}, \text{ or}$$

$$\lambda < \frac{t/0.25}{0.25+t} = \frac{16t}{1+4t} \text{ per hour.}$$

Note that if $t$ = 1/12 h (5 min), then this formula yields λ < 1, which matches our previous calculation. If we plot the maximum arrival rate (λ) as a function of the allowable waiting time ($t$) for the Macrohard service system, we get the curve in Fig. 4.5. Notice that in order for the system to handle 3.5 customers per hour, which is still well below the capacity of 4 per hour, we must accept an average waiting time of 2 h! Clearly, the presence of variability can cause a great deal of congestion in queueing systems, which makes it difficult to operate them near their capacity.

This insight is behind the fact that semiconductor facilities (wafer fabs) typically aim for equipment utilization of around 75%. If they are operated at higher levels of utilization, wafers wait too long at individual processing stations and hence the cycle time (i.e., time to produce a wafer) becomes uncompetitively long.

There are still two questions about the Macrohard call center that we have yet to address: What factors affect the likelihood and duration of customer waiting? and What options exist for reducing customer waiting time? To give useful answers to these, and to generalize the previous answers to situations where we cannot assume that arrivals and services are Markov, we need more powerful tools.

## M/G/1 Queue

The M/M/1 queueing model gives us some important insights. In particular, it shows that variability causes waiting and that waiting increases with utilization. But because it assumes that both interarrival times and service times are Markov (exponentially distributed), there is no way to adjust the amount of variability in an M/M/1 system. To really understand how variability drives the performance of a queueing system, we need a more general model. The simplest option is the so-called M/G/1 queue, in which interarrival times are still Markov but service times, are general (represented by the G in the label). This means that service times can take on any probability distribution, as long as they are independent of one another.

To express and examine the M/G/1 queueing model, we need to characterize the variability of the service times. The most common measure of variability used in probability and statistics writings is the **standard deviation**, which we usually denote by σ (sigma). While this is a perfectly valid measure of variability, since it measures the "spread" of a probability distribution, it does not completely characterize variability of a random variable. To see why, suppose we are told that the standard deviation of a service time is 5 min. We cannot tell whether this constitutes a small or large amount of variability until we know the mean service time. For a system with a mean service time of 2 h, a 5 min standard deviation does not represent much variability. However, for a system with a 2 min average service time, a 5 min standard deviation represents a great deal of variability.

To provide a more general measure of variability, we define the **coefficient of variation (CV)** of a random variable to be the standard deviation divided by the mean. So, the **service time CV**, denoted by *cs*, is given by

$$c_s = \frac{\sigma}{\tau}, \tag{9}$$

where σ represents the standard deviation of the service time and τ represents the mean service time.

Figure 4.6 illustrates three distributions of service time with a mean of 15 min but with CVs of 0.33, 1, and 2. The distribution with $c_s = 0.33$ has a normal shape, with service times symmetrically distributed around the mean. This represents a case with a fairly low level of variation in service times. In contrast, the distribution with $c_s = 2$ exhibits a high frequency of very short service times (i.e., less than 5 min), but also a non-negligible frequency of very long service times (i.e., over 60 min). Hence, this distribution represents a high level of variability. The distribution with $c_s = 1$ lies between these two cases, with more spread than the $c_s = 0.33$ case but less spread than the $c_s = 2$ case. In fact, the case with $c_s = 1$ is precisely the **exponential distribution**, which was illustrated in Fig. 4.1.

A host of factors can influence the value of $c_s$ in real-world queueing systems. In the Macrohard example, $c_s$ will be large if many callers have simple questions

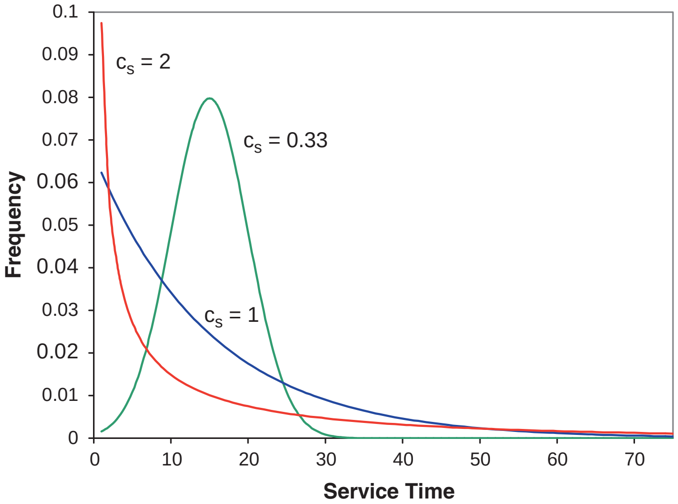


**Fig. 4.6** Service time distributions with various coefficients of variation

that can be answered quickly, but a few have complicated problems that take a long time to resolve. If this were the case, then the 15 min average service time could consist of a number of 2–3-min calls interspersed with an occasional 1 h call. As we will see below, this leads to very different behavior than the situation where the 15 min average is made up of service times that are clustered around 15 min (e.g., as in the $c_s = 0.33$ case in Fig. 4.6).

Because the standard deviation is always equal to the mean in the exponential distribution, $c_s = 1$ in the M/M/1 queue. Since we cannot change $c_s$ in the M/M/1 model, we cannot use it to examine the effect of changes in service time variability. In the M/G/1 queue, however, $c_s$ can be anything from zero (which indicates deterministic service times with no variability at all) on up. This allows us to adjust variability and examine the impact on system behavior.

Deriving the expression for mean waiting time in the M/G/1 cannot be done using the birth–death approach used for the M/M/1 case. The reason is that since service times are no longer memoryless, knowing the number of customers in the system is not sufficient to predict future behavior. We must also know how long the current customer (if any) has been in service. This requires a more sophisticated analysis approach, which is beyond the scope of this chapter.

So instead we simply state the result for the mean waiting time (i.e., queue time) for the M/G/1 queue, which is

$$W_q^{M/G/1} = \left(\frac{1+c_s^2}{2}\right)\left(\frac{\rho}{1-\rho}\right)\tau. \tag{10}$$

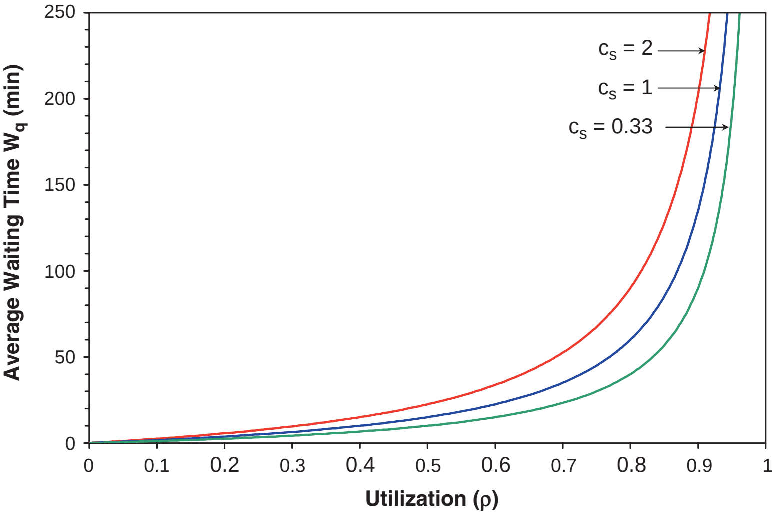


**Fig. 4.7** Effect of variability and utilization on average waiting time (τ = 15 min)

The total mean time in the system is hence given by

$$W^{M/G/1} = \left(\frac{1+c_s^2}{2}\right)\left(\frac{\rho}{1-\rho}\right)\tau + \tau. \tag{11}$$

Equation (10) is known as the **Pollaczek-Khintchine (P-K) formula**. The only difference between this equation and the corresponding formula (Eq. (8)) for the M/M/1 queue is the presence of the term $(1 + c_s^2)/2$. If $c_s = 1$, then this term becomes 1 and Eq. (10) reduces to Eq. (8), as it should. But if $c_s > 1$ then the waiting time in queue will be larger in the M/G/1 queue than in the M/M/1 queue. Likewise, if $c_s < 1$ then the waiting time will be smaller in the M/G/1 queue than in the M/M/1 queue.

Figure 4.7 illustrates the P-K formula for the Macrohard service support system with an average service time of τ = 15 min and service time CV's of $c_s = 0.33$, 1 and 2. As we already know, increasing $c_s$ increases the average waiting time $W_q$. More interestingly, Fig. 4.7 shows that the amount by which average waiting increases due to an increase in service time CV increases rapidly as server utilization increases. For instance, at a utilization of ρ = 0.1, increasing $c_s$ from 0.33 to 2 increases the waiting time by about a minute and a half. At a utilization of ρ = 0.9, the same increase in $c_s$ increases waiting time by over 100 min. The reason for this is that, as the P-K formula (Eq. (10)) shows, the variability term $(1+ c_s^2)/2$ and the utilization term ρ/(1-ρ) are multiplicative. Since the utilization term increases exponentially as ρ approaches one, variability in service times is most harmful to performance when the system is busy.

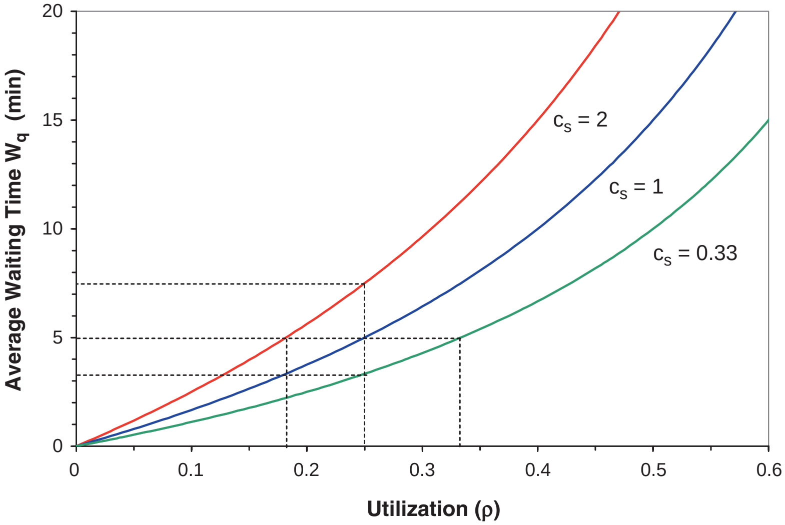


**Fig. 4.8** Tradeoffs between variability reduction, waiting time, and utilization

To appreciate the practical significance of the combined effect of utilization and variability in causing waiting, we zoom in on the curves of Fig. 4.7 and display the result in Fig. 4.8. For the case where $c_s = 1$ (i.e., the M/M/1 queue), we see that a 5 min average waiting time requires a utilization level of 0.25, as we computed earlier. Now suppose we were to reduce $c_s$ to 0.33. In the Macrohard system this might be achievable through technician training or development of an efficient diagnostic database. By reducing or eliminating excessively long calls, these measures would lower the standard deviation, and hence the CV, of the service times. (Shortening the time to handle long calls without lengthening the time to handle other calls would also reduce the average service time, which would further improve performance. But for clarity, we ignore this capacity effect and focus only on variability reduction.)

Figure 4.8 shows that reducing the service time variability to $c_s = 0.33$ reduces the average waiting time from 5 min to a bit over 3 min (3 min and 20 s to be exact).

Another way to view the impact of reducing $c_s$ from 1 to 0.33 is to observe the change in the maximum utilization consistent with an average waiting time of 5 min. Figure 4.8 shows that the Macrohard service center with $c_s = 0.33$ can keep average waiting time below 5 min as long as utilization is below 0.33 (This is coincidence that utilization is the same as $c_s$). Compared with the utilization of 0.25 that was necessary for the case with $c_s = 1$, this represents a 32% increase. Hence, in a very real sense, variability reduction acts as a substitute for a capacity increase, which can be an enormously useful insight.

For example, suppose Macrohard is currently experiencing a call volume of 1 per hour and is just barely meeting the 5-minute constraint on customer waiting time. Now suppose that, due to increased sales, the call volume goes up by 15%. If Macrohard does not want average waiting time to increase and cause customer discontent, it must do something. The obvious, but costly, option would be to add a second service technician. But, as our discussion above shows, an alternative would be to take steps to reduce the variability of the service times. Doing this via training or technology (e.g., a diagnostic database) might be significantly cheaper than hiring another technician.

Production systems represent another environment where this insight into the relationship between variability and capacity in queueing systems is important. In such systems, each workstation represents a queueing system. The time to process jobs (e.g., machine parts, assemble products, or complete other operations) includes both actual processing time and various types of "nonvalue-added" time (e.g., machine failures, setup times, time to correct quality problems, etc.). These nonvalue-added factors can substantially inflate the process time CV at workstations. Hence, ameliorating them can reduce cycle time (i.e., total time required to make a product), increase throughput (i.e., production rate that is consistent with a cycle time constraint) or reduce cost by eliminating the need to invest in additional equipment and/or labor to achieve a target throughput or cycle time. The techniques used to reduce variability in production systems are often summarized under the headings of **lean production** and **six sigma quality control.** Hence, whether most practitioners know it or not, the science of queueing lies at the core of much of modern operations management practice.

## G/G/1 Queue

The M/G/1 model gives us insight into the impact of service time variability on tradeoff between average waiting time and customer volume. But because interarrival times are still assumed to be Markov (memoryless), we cannot alter arrival variability. So we cannot use this model to examine policies that affect arrivals into a queueing system.

For example, suppose Macrohard were to adopt a scheduling system for appointments with technicians. That is, instead of simply calling in at random times, customers seeking help must visit the Macrohard website and sign up for a phone appointment. If appointments are $x$ minutes apart (and all slots are filled) then calls will come in at a steady rate of one every $x$ minutes. This (completely predictable) arrival process is vastly different from the Markov arrival process assumed in the M/M/1 and M/G/1 models.

Figure 4.9 graphically illustrates the even spacing of deterministic (scheduled) arrivals and the uneven spacing of (random) Markov arrivals. Because interarrival times are highly variable in a Markov arrival process, customers tend to arrive in clumps. These clumps will cause periods of customer waiting and hence will tend to inflate average waiting time. To see how much, we need to extend the M/G/1 queueing model to allow different levels of arrival variability.

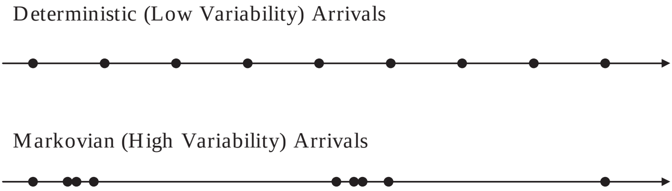

Deterministic (Low Variability) Arrivals
Markovian (High Variability) Arrivals

**Fig. 4.9** Low and high variability arrival processes

To develop such a model, we characterize arrival variability in exactly the same way we characterized service time variability in Eq. (9). That is, we define the **interarrival time CV** by

$$c_a = \frac{\sigma_a}{\tau_a}, \tag{12}$$

where $\tau_a$ represents the mean time between arrivals (and so $\lambda = 1/\tau_a$ is the average arrival rate) and $\sigma_a$ represents the standard deviation of interarrival times. A higher value of $c_a$ indicates more variability or "burstiness" in the arrival stream. For instance, the constant (deterministic) arrival stream in Fig. 4.9 has $c_a = 0$, while the Markov arrival stream has $c_a = 1$.

The parameter $c_a$ characterizes variability in interarrival *times*. But we can also think of this variability in terms of arrival *rates*. That is, the arrival bursts in the high variability case in Fig. 4.9 can be viewed as intervals where the arrival rate is high, while the lulls represent low arrival rate intervals. Examples of variable arrival rates are common. Fast food restaurants experience demand spikes during the lunch hour. Toll booths experience arrival bursts during rush hour. Merchandising call centers experience increases in call volume during the airing of their television "infomercials."

The impact of increased fluctuations in arrival rate is similar to that of increased arrival variability (i.e., increased $c_a$)−both make it more likely that customers will arrive in bunches and hence cause congestion. However, an important difference is that arrival fluctuations are often predictable (e.g., we know when the lunch hour is) and so compensating actions (e.g., changes in staffing levels) can be implemented.

Throughout this chapter we assume that variability, in interarrival times and service times, is unpredictable. In the Macrohard example this is a reasonable assumption over small intervals of time, say an hour, since we cannot predict when customers will call in or how long they will require to service. However, some variability over the course of a day may be partially predictable. For instance, we may know from experience that the call rate is higher on average from 2–3 PM than from 2–3 AM. As an approximation, we could simply apply our queueing models using a different arrival rate (λ) for each time interval. For more detailed analysis of the behavior of queueing systems with nonstationary arrivals, practitioners often make use of discrete event simulation.

Using $c_a$ as a measure of arrival variability we can approximate the average waiting time of the G/G/1 queue as follows:

$$W_q^{G/G/1} \approx \left(\frac{c_a^2 + c_s^2}{2}\right)\left(\frac{\rho}{1-\rho}\right)\tau. \tag{13}$$

Hence the total system time is given by

$$W^{G/G/1} \approx \left(\frac{c_a^2 + c_s^2}{2}\right)\left(\frac{\rho}{1-\rho}\right)\tau + \tau. \tag{14}$$

Equation (13) is sometimes called **Kingman's equation** after one of the first queueing researchers to propose it (Kingman, 1966). Note that it differs from Eq. (10) for the M/G/1 queue only by replacing a "1" by $c_a^2$. Hence, when arrivals are Markov, and so $c_a = 1$, Eq. (13) reduces to Eq. (10) and gives the exact waiting time for the M/G/1 queue. For all other values of $c_a$, Eq. (13) gives an approximation of the waiting time.

Because it is so similar to the P-K formula, the behavior of the G/G/1 queue implied by Kingman's equation is very similar to the behavior of the M/G/1 queue we examined previously. Specifically, since the utilization term is the same in Eqs. (10) and (13), the average waiting time in the G/G/1 queue ($W_q^{G/G/1}$) also increase exponentially as utilization (ρ) approaches one, as illustrated for $W_q^{G/G/1}$ in Fig. 4.7. The only difference is that, in the G/G/1 case, increasing either arrival variability ($c_a$) or service variability ($c_s$) will increase waiting time and hence cause $W_q^{G/G/1}$ to diverge to infinity more quickly. An interesting and useful insight from Kingman's equation is that arrival variability and service variability are equally important in causing waiting.

As we observed from the P-K formula, Kingman's equation implies that variability has a larger impact on waiting in high utilization systems than in low utilization systems. Mathematically the reason for this is simple: the variability ($[c_a^2 + c_a^2]/2$) and the utilization ($\rho/[1-\rho]$) terms are multiplicative. Hence, the larger the utilization, the more it amplifies the waiting caused by variability.

While this mathematical explanation of the interaction between variability and utilization is neat and clean, it is less than satisfying intuitively. To gain insight into how variability causes congestion and why utilization magnifies it, it is instructive to think about the evolution of the number of customers in the system over time.

Figure 4.10 shows sample plots of the number in system for two single server queueing systems that have the same utilization (i.e., same fraction of time the server is busy). Figure 4.10(a) represents a low variability system, in which customers arrive at fairly regular intervals and service times are also fairly constant. As a result, each customer who arrives to this system finds it empty and the number of customers never exceeds one. Thus, no one experiences a wait. In Kingman equation terms,

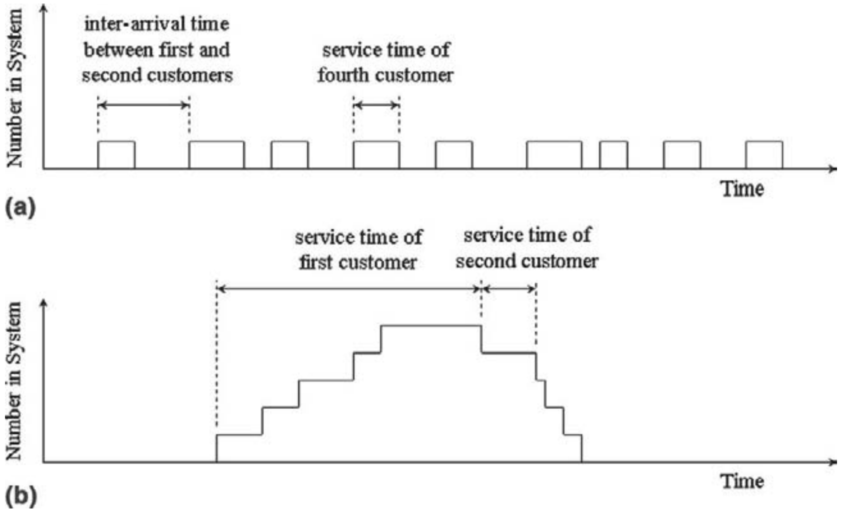


**Fig. 4.10** Number of customers in (**a**) low variability system and (**b**) high variability system

Fig. 4.10(a) represents a system where $c_a$ and $c_s$ are both low and hence so is average waiting time.

Figure 4.10(b) shows the very different behavior of a high variability system. Arrivals (instants where the number in system increases by one) occur in a bunch, rather than spread over the entire interval as in Fig. 4.10(a). Furthermore, the service times are highly variable; the first service time is very long, while all subsequent ones are quite short. As a result, all customers after the first one wind up waiting for a significant time in the queue. In Kingman equation terms, Fig. 10(b) represents a system where $c_a$ and $c_s$ are both high and hence average waiting time is also large.

To understand why variability causes more congestion and waiting when utilization is high than when utilization is low we examine the impact of an additional arrival on the average waiting time. Figure 4.11(a) shows the number of customers in a low utilization system. Figure 4.11(b) shows the modified figure when an extra arrival is introduced. Since this arrival happened to find the server idle (as is likely when system utilization is low) and was completed before the next arrival, this added arrival did not increase average waiting time. The lower the utilization, the more likely an additional arrival will be processed without waiting. Even if the arrival were to occur during a busy period, the extra waiting would be slight because another idle interval will occur before too long. Hence, when utilization is low we can conclude: (a) average waiting time will be low, and (b) a slight increase in utilization will increase waiting time only modestly.

In contrast, Fig. 4.12 illustrates what happens when an additional arrival is added to a high utilization system. As shown in Fig. 4.12(a), a high utilization has few and short idle periods. Hence, a randomly timed arrival is likely to occur when

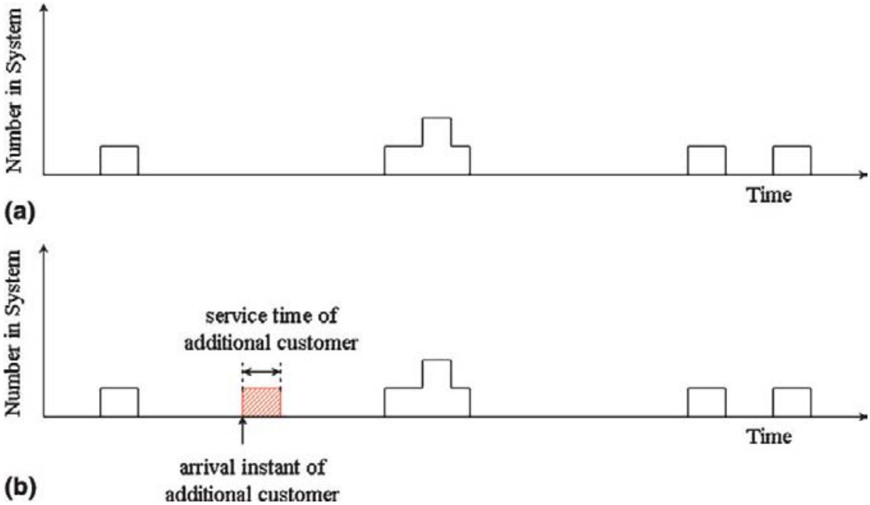


**Fig. 4.11** (**a**) Low utilization system and (**b**) low utilization system with additional customer

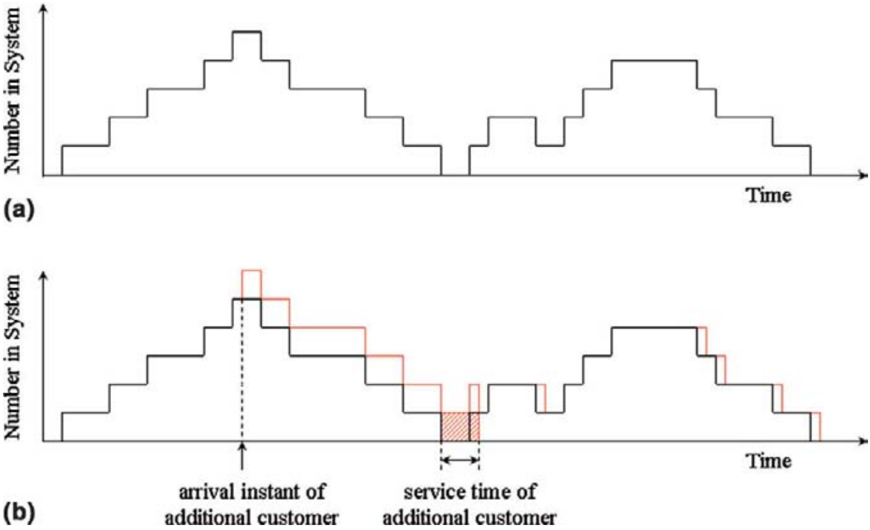


**Fig. 4.12** (**a**) High utilization system and (**b**) high utilization system with additional customer

the server is busy and customers are waiting. When this occurs, as shown in Fig. 4.12(b), the number in system will increase by one from the time of the new arrival until the end of the busy period during which it arrived. It will also extend this busy period by the service time of the new arrival. This could also cause the current busy period to run into the next one, causing additional waiting and a lengthening of that busy period as well. Only when the extra service time represented by the new arrival is "absorbed" by idle periods will the effect on waiting dissipate. Consequently, the busier the system, the less available idle time and hence the longer the effect of a new arrival on the number in system will persist. Hence, when utilization is high, we conclude: (a) the impact of variability on waiting will be high, and (b) a slight increase in utilization can significantly increase average waiting time.

Finally, armed with Kingman's equation and the intuition that follows from it, we return to the Macrohard case to address the remaining two questions.

*What factors affect the likelihood and duration of customer waiting?* Kingman's equation provides a crisp summary of the drivers of waiting time: arrival variability ($c_a$), service variability ($c_s$) and utilization ($\rho$), which is composed of the arrival rate ($\lambda$) and the average service time ($\tau$). By exploring the underlying causes of these four parameters ($c_a$, $c_s$, $\lambda$, $\tau$), we can systematically identify the factors that cause waiting in a given queueing system. We do this for the Macrohard case in order to answer the last question we posed concerning its performance.

*What options exist for reducing customer waiting time?* Because the utilization term frequently dominates Kingman's equation, we first discuss options for reducing utilization (i.e., reducing $\lambda$ or $\tau$) and then turn to options for reducing variability.

1. *Reduce arrival rate ($\lambda$):* There are many ways to reduce the call rate to the Macrohard call center. However, some of these (e.g., making it inconvenient for customers to get through) are in conflict with the strategic objective of providing good service. But there are options that could both reduce the call rate and increase customer satisfaction. For example, improving user documentation, providing better web-based support, making user interfaces more intuitive, offering training, etc., are examples of actions that may make customers less reliant on technical support.
2. *Reduce service time ($\tau$):* Similarly, there are options for reducing service times (e.g., hanging up on customers or giving them curt answers) that are inconsistent with Macrohard's strategic objectives. But options such as hiring more qualified technicians, improving technician training, and providing technicians with easy-to-use diagnostic tools, could both speed and improve customer service. Policies, such as better self-help tools, that enable customers to partially solve problems and collect information of use to the technician prior to the service call could also speed service. However, it is worth asking whether customers will be happier performing such self-help rather than calling for technical support immediately. Finally, it may be possible for the technician to multi-task, by putting one customer who has been asked to perform some diagnostic steps on hold and speaking with another customer. If processing multiple customers at once services them more quickly than processing them sequentially (and doesn't anger them with more hold time) then this could be an option for dealing with bursts of calls.
3. *Reduce service variability ($c_s$):* The options for reducing service time may reduce service time variability as well. For example, a better trained technician is less likely to "get stuck" on a customer problem and will therefore have fewer exceptionally long calls. Eliminating very long calls will reduce both the mean and the

variability of service times. Similarly, options that promote customer self-help may shorten the longest calls and thereby reduce service time variability. Finally, while Macrohard does not have this option on account of their system consisting of a single technician, larger call centers can reduce variability by stratifying calls. For instance, most companies have a voice menu to separate customers with (short) sales calls from customers with (long) technical support calls. In addition to reducing variability in service times, such a policy allows service agents to specialize on a particular type of call and, we hope, provide better service.

4. *Reduce arrival variability ($c_a$):* We noted earlier that interarrival times tend to be Markov (i.e., $c_a = 1$) when arrivals are the result of independent decisions by many customers. In most call centers, including that of Macrohard, it is reasonable to assume that this will be the case, provided that customers call in whenever they like. But it isn't absolutely essential that customers be allowed to call in at will. Many service systems (e.g., dentists, car mechanics, college professors) require, or at least encourage, customers to make appointments.

Suppose that customers were required to schedule technical support appointments in the Macrohard support system by going to a website and booking a call-in time. Recall that we used the M/M/1 model to find that if Macrohard wishes to keep average waiting time below 5 min when average process times are 15 min, the arrival rate can be no higher than one call per hour. This calculation, however was based on an assumption of Markov arrivals and service times ($c_a = c_s = 1$). If we do not alter the variability in service times (so $c_s = 1$) but we schedule arrivals (so $c_a = 0$), how far apart must we schedule appointments to ensure that average waiting times remain below 5 min?

Recalling that $\rho = \lambda\tau$, and expressing all times in units of hours, we can use Eq. (13) (Kingman's equation) to answer this question as follows:

$$W_q^{G/G/1} \approx \left(\frac{c_a^2 + c_s^2}{2}\right)\left(\frac{\lambda\tau}{1-\lambda\tau}\right)\tau = \left(\frac{0^2+1^2}{2}\right)\left(\frac{\lambda(0.25)}{1-\lambda(0.25)}\right)0.25 < \frac{5}{60}, \text{ or}$$

$$\left(\frac{\lambda(0.25)}{1-\lambda(0.25)}\right) < \frac{40}{60}, \text{ or}$$

$$\lambda < 1.6 \text{ per hour.}$$

This shows that smoothing out arrivals to the support center by asking customers to sign up for times will allow the technician to handle 60% more calls (i.e., λ increases from 1 per hour to 1.6 per hour) with the same average waiting time of 5 min. The reason, as we noted above, is that the technician will not have to cope with clumps of calls that cause occasional intervals of congestion. So the technician can handle more calls if they come in more smoothly.

Of course, customers may feel that being forced to make appointments to get technical support is not consistent with good service. So Macrohard may elect to continue allowing customers to call whenever they like. But they will need a larger capacity (i.e., more and/or faster technicians) to accommodate unscheduled customers without long delays.

Clearly, there are systems in which long delays and/or low server utilization are worse than the inconvenience of making appointments. We schedule appointments with physicians, hair stylists, dance instructors and many other service providers. If instead we simply showed up at our leisure and requested service we would frequently experience long waits (unless the provider had very low utilization, which would probably not be a good sign with respect to quality).

Finally, in addition to the four categories of improvement suggested by Kingman's equation, Macrohard could also consider some structural approaches for reducing average waiting time. One possibility is customer sequencing. As noted in Chap. 1, the shortest processing time (SPT) rule minimizes average flow time for a fixed set of jobs. While queueing systems do not match the conditions under which the SPT rule is strictly optimal (i.e., because arrivals are dynamic), they share some behavior in common with single station scheduling systems. If Macrohard can identify customers (e.g., using a voice menu system) who require short service times, they can very likely reduce average waiting time by handling them first and then taking the customers with long complex questions.

Another improvement approach might be to convert the system to a priority queue. If some customers seeking service are more important than others (e.g., have paid for premier support or have indicated through the phone menu that they have more pressing problems), then dividing customers into separate queues and handling the higher priority customers first could reduce the waiting time of the most important customers. Of course, this improvement would come at the expense of increased waiting time for lower priority customers. But such a shift might be consistent with Macrohard's customer support strategy.

Finally, Macrohard could take steps to shift arrivals from busy (high utilization) periods to periods with excess capacity. One way to do this would be to allow only a fixed number of calls in queue at any time. If a customer calls when this buffer is full, then he/she would get a busy signal (or more appropriately a polite message) and would have to call back later. Alternatively, Macrohard could implement a system to announce the expected waiting times and count on customers to balk on joining the queue when the system is too busy.

## Applications

Although we tend to think of queueing primarily in the context of service systems (e.g., waiting in line at the bank, grocery store or movie theater), queueing behavior occurs in all kinds of systems.

In manufacturing plants, jobs await processing at work stations. Variability in process times can be the result of product variety, machine failures, setup times, quality problems, operator variability, and many other causes. Variability in interarrival times can be the result of uneven customer orders, upstream process variability, material handling, etc. High utilization work stations are called “bottlenecks” and are responsible for inflating work-in-process (WIP) and cycle times. Lean manufacturing methods focus on supplementing bottleneck capacity (e.g., through cross-training workers to “float” between work stations as needed) and reducing variability (e.g., through total preventive maintenance and setup reduction).

In transportation systems, queueing-type congestion occurs at many locations, including toll booths, stop signs, entrance ramps, etc. As we all know, when traffic gets heavier (utilization increases) congestion grows. Traffic calming methods, such as lights that pace cars on highway entrance ramps, reduce congestion by reducing variability. Electronic tags for toll collection (E-ZPass, I-Pass, SunPass, etc.) both increase capacity and reduce service variability at toll booths. While models of vehicle flow go well beyond single server queues, the basic insights of this chapter are at the core of traffic science.

Queuing behavior abounds in telecommunications systems. Calls at exchanges, information packets at routers and jobs at central processing units (CPUs) are just a few examples of electronic queueing systems. A prevalent application of models like those discussed in this chapter is sizing web server capacity. Since arrivals to a web site (hits) are random and downloaded file sizes are variable, queues will form. To prevent slow response, the web site must have sufficient server capacity to keep average utilization to reasonable levels.

An application where rapid response is essential is emergency services. Police, fire, and ambulance services can all be viewed as queueing systems with calls constituting arrivals. Since excessive waiting time is not acceptable in these systems and variability is unavoidable (people do not schedule their heart attacks), the only way to keep waiting to a minimum is to have a great deal of excess capacity. As a result, ambulance systems operate at only 10–15% utilization and fire fighters tend to spend a lot of time cleaning their equipment while they are waiting for a call. But when one comes, they are likely to be ready to go.

These are just a few of the many situations where queueing behavior arises. Although the details differ widely, the main insights are remarkably similar.

## Summary of Insights

The single server queueing models considered in this chapter are simple, but they illustrate essential behaviors that extend well beyond these simple settings. These are:

1. *Variability causes congestion:* Uneven interarrival or service times in a queueing system result in occasional backups during which customers must wait. The more frequent and more severe these backups, the larger the expected wait time will be. Kingman’s Equation shows that interarrival and service variability (as measured by coefficient of variation) play equal roles in causing congestion and waiting.

This insight explains why we wait at both doctors' offices and ATM's. Since patients have appointments with doctors, arrivals are quite steady and predictable (i.e., $c_a$ is low). So it is variability in the amount of time the doctor spends with individual patients that causes the queueing that makes us wait. In contrast, arrivals to an ATM are not scheduled and are therefore quite random, while service times are fairly uniform. Hence, it is arrival variability ($c_a$) that causes us to wait at an ATM.

2. *Utilization exacerbates congestion caused by variability:* The busier a server is, the more vulnerable it is to the backups that result from uneven interarrival or service times. A key feature of Kingman's Equation is that the variability term $((c_a^2 + c_a^2)/2)$ and the utilization term $(\rho/(1\text{-}\rho))$ are multiplicative. Doubling the variability term in a system with $\rho = 0.5$ will double waiting time in queue, while doubling the variability term in a system with $\rho = 0.9$ will increase waiting time by 18 times. Clearly, variability has an extreme effect on busy systems.

This insight explains the very different priorities of emergency medical service (EMS) systems and call centers. Because they must be highly responsive, EMS systems operate at very low utilization levels (in the range of 10–15%). Hence, a change in the variability of response times ($c_s$) will not have a large impact on average waiting time and so service time smoothing is not an attractive improvement policy in EMS systems. In contrast, call centers operate at fairly high utilization levels (80% or more) in pursuit of cost efficiency. This makes them sensitive to variability in service times and so training, technology and other policies that facilitate uniform response times can be of substantial value.

3. *Service and production systems with variability cannot operate at their theoretical capacity:* All of the formulas (Eqs. (8), (10), and (13)) for average waiting time in a single server queue have a $(1\text{-}\rho)$ term in the denominator, which implies that waiting time will approach infinity as $\rho$ approaches one. Of course, the queue length can only grow to infinity given an infinite amount of time. So what this really means is that systems with 100% utilization will be highly unstable, exhibiting a tendency for the queue length and waiting time to get out of control. Hence, over the long term, it is impossible to operate at full utilization. Indeed, in systems with significant variability and a limit on the allowable waiting time, the maximum practical utilization may be well below 100%.

This insight may seem at odds with every day behavior. For instance, plant managers are fond of saying that they operate at more than 100% capacity. That is, their plant is rated at 2,000 circuit boards per day but they have been producing 2,500 per day for the past month. While it is possible to temporarily run a queueing system at or above capacity, this cannot be done indefinitely without having congestion increase to intolerable levels. So what the plant managers who make statements like this usually mean is that they are running above the capacity defined for regular operating conditions. But they typically do this by using overtime, scheduling more operators, or taking other measures that increase their effective capacity. With capacity properly redefined to consider these increases, utilization *will* be below 100% over the long term.

4. *Variability reduction is a substitute for capacity:* Since both variability and utilization cause congestion, policies that address either can be used to reduce waiting. Conversely, if the maximum waiting time is constrained by strategic considerations, then either reducing variability or increasing capacity can increase the volume the system can handle.

The Macrohard example illustrates this insight by showing that reducing variability by scheduling support calls will allow the technician to handle 60% more calls than if customers call in at random. While the theoretical capacity of the technician is not changed by scheduling calls, the practical capacity of the system (i.e., the call rate that can be handled without allowing average waiting time to grow beyond 5 min) is increased due to the reduction in congestion.

This insight has powerful implications in manufacturing, as well as service systems. For instance, in a batch chemical process, in which reactors are separated by tanks that hold small amounts of material, the work in process (and hence the time to go through the line) is strictly limited. Since reactors, which act like single server queueing systems, cannot build up long queues due to the limited storage space, variability prevents them from achieving high levels of utilization. Specifically, the reactors will be *blocked* by a full downstream tank or *starved* by an empty upstream tank. The more variability there is in processing times, the more blocking and starving the reactors will experience. Therefore, if changes in reaction chemistry, process control or operating policies can make processing times more regular, the line will be able to operate closer to its theoretical capacity. This would allow use of a smaller line to achieve a target throughput and hence variability reduction is indeed a substitute for capacity.

Insights like these will not eliminate waiting in our daily lives. But they can help us understand why it occurs and where to expect it. More importantly, in the hands of the right people, they can facilitate design of effective production and service systems. As such, the science of queueing offers the dream of a world in which waiting is part of sensibly managed tradeoffs, rather than an annoying consequence of chaotic and ill-planned decisions.

## Historical Background

The scientific study of queues began nearly 100 years ago with the work of A.K. (Agner Krarup) Erlang, an engineer working for the Copenhagen Telephone Company. Motivated by the problem of determining how many circuits were required to provide acceptable telephone service, Erlang made use of the Poisson distribution to model randomness in phone calls (Erlang 1909) and developed formulas for average waiting time and loss rates for calls (Erlang 1917).

Although they were not yet labeled as such, the queueing models used by Erlang were relatives of the M/M/1 model discussed in this chapter. Over the next half century, researchers extended their analysis to include the M/G/1 queue

(Khintchine 1932), the M/G/C (Pollaczek 1934) and G/G/1 queue (Lindley 1952) and their many variants. However, while these early papers talked about congestion and used the term "queue," the term "queueing" was not coined until after World War II by David Kendall (1951). Kendall also systematized the field he named by introducing the standard A/B/C classification in which A designates the arrival process, B represents the service process and C denotes the number of servers (Kendall 1953). In this chapter we have only considered as values for A and B the options of M (Markov) and G (general). But many others, such as D (deterministic), Ek (Erlang), and Ph (phase type), are possible, which makes the study of queues incredibly rich.

Stimulated by the post-war burst of activity by operations research scholars to apply mathematics to practical problems, queueing theory enjoyed a period of substantial growth during the 1960s and 1970s. Indeed, by 1986, the literature on queues had become so extensive that a new journal, Queueing Systems, was devoted exclusively to queueing research. The rich history of the mathematics of queues is far too extensive to document here, but Stidham (2002) offers an excellent review.

From an application standpoint, two developments—approximations and networks—were particularly important. The work of Kingman (1966), based on heavy traffic analysis, led to the approximation for the G/G/1 queue discussed in this chapter. Because this approximation, and similar ones for variants of the G/G/1, depend only on two moments (mean and coefficient of variation) of the interarrival and service times, they are enormously useful in modeling systems with limited data and drawing general insights about the causes of congestion.

Jackson (1957, 1963) initiated the study of networks of queues, in which entities flow through a series of servers and may have to wait at any of them. Because systems in manufacturing, service, transportation, telecommunications and many other domains are often characterized as flow networks, queueing network models are highly useful in analyzing and improving the performance of such systems. Whitt (1983) combined heavy traffic approximations with network analysis to produce the first practical tool for analyzing networks of queues. Today, there are a variety of commercial software packages that rely on queueing network models to make performance analyses of various operations systems.

The progress on the mathematics of queues has facilitated a broad range of applications of queueing. In the 1970s, researchers (e.g., Larson 1972; Larson and Odoni 1981) made heavy use of queueing models to make quantitative analyses of urban systems. Also in the 1970's, queueing models became a standard tool in the performance analysis of computer systems (see Kleinrock 2002 for a historical overview). The 1980s saw the adoption of queueing methods in transportation systems analysis (Newell 1982). In the 1990s, a number of scholars (e.g., Buzacott and Shantikumar 1993; Hopp and Spearman 2000) relied on queues to construct a science of manufacturing that underlies the practices of lean production, agile manufacturing and six sigma. More recently, concerns about security have motivated scholars (e.g., Wein et al. 2003; Green and Kolesar 2004) to use queueing models for improving emergency response systems.

The past century has seen remarkable advances in our understanding of the fundamental causes of waiting. But since waiting remains a major source of inefficiency in many public and private sector systems, queueing theory and its applications will continue to receive research attention for many years to come.

## References

Buzacott, J. and J.G. Shantikumar. (1993) *Stochastic Models of Manufacturing Systems*. Prentice-Hall, Englewood Cliffs, NJ.

Erlang, A.K. (1909) “The theory of probabilities and telephone conversations,” *Nyt Tidsskrift for Matematik B* **20**, 33–39.

Erlang, A.K. (1917) “Solution of some problems in the theory of probabilities of significance in automatic telephone exchanges,” *Elektrotkeknikeren* **13**, 5–13.

Green, L.V. and P.J. Kolesar. (2004) “Applying management science to emergency response systems: Lessons from the past,” *Management Science* **50**(8), 1001–1014.

Hopp, W. and M. Spearman. (2000) *Factory Physics: Foundations of Manufacturing Management*. McGraw-Hill, New York.

Jackson, J.R. (1957) “Networks of waiting lines,” *Operations Research* **5**, 518–521.

Jackson, J.R. (1963) “Jobshop-like queueing systems,” *Management Science* **10**, 131–142.

Kendall, D.G. (1951) “Some problems in the theory of queues,” *Journal of the Royal Statistical Society B* **13**, 151–185.

Kendall, D.G. (1953) “Stochastic processes occurring in the theory of queues and their analysis by the method of the embedded Markov chain,” *Annals of Mathematical Statistics* **24**, 338–354.

Khintchine, A. (1932) “Mathematical theory of a stationary queue,” *Mathematicheskii Schornik* **39**, 73–84.

Kingman, J.F.C. (1966) “On the algebra of queues,” *Journal of Applied Probability* **3**, 285–326.

Kleinrock, L. (2002) “Creating a mathematical theory of computer networks,” *Operations Research* **50**, 125–131.

Larson, R. (1972) *Urban Police Patrol Analysis.* MIT Press, Cambridge, MA.

Larson, R. and A. Odoni. (1981) *Urban Operations Research*. Prentice-Hall, New York.

Lindley, D.V. (1952). The theory of queues with a single server. *Proceedings of the Cambridge Philosopical Society* **48**, 277–289.

Newell, G.F. (1982). *Applications of Queueing Theory*. 2nd edn., Chapman & Hall, New York.

Pollaczek, F. (1934). Uber das waterproblem. *Mathematische Zeitschrift* **32**, 492–537.

Stidham, S. (2002). Analysis, design, and control of queueing systems. *Operations Research* **50**(1), 197–216.

Wein, L.M., D.L. Craft and E.H. Kaplan. (2003). Emergency response to an anthrax attack. *Proceedings of the National Academy of Science* **100**(7), 4346–4351.

Whitt, W. (1983). The queueing network analyzer. *Bell System Technical Journal* **62**, 2779–2815.

# Chapter 5
# Little's Law

**John D.C. Little and Stephen C. Graves**
**Massachusetts Institute of Technology**

*The average waiting time and the average number of items waiting for a service in a service system are important measurements for a manager. Little's Law relates these two metrics via the average rate of arrivals to the system. This fundamental law has found numerous uses in operations management and managerial decision making.*

## Introduction

Caroline is a wine buff and bon vivant. She likes to stop at her local wine store, *Transcendental Tastings*, on the way home from work. She browses the aisles looking for the latest releases from her favorite vineyards. Occasionally she picks up a few bottles. She stores these in a rack in a cool corner of her cellar. She and her partner eat out frequently but when they are at home they usually split a bottle of wine at dinner. Sometimes they have friends over and that puts a bigger dent in the wine inventory.

They have been doing this for some time. Her wine rack holds 240 bottles. She notices that she seldom fills the rack to the top but sometimes after a good party the rack is empty. On average it seems to be about 2/3rds full, which would equate to 160 bottles.

Many wines improve with age. After reading an article about this, Caroline starts to wonder how long, on average, she has been keeping her wines. She went back through a few months of wine invoices from *Transcendental* and estimates that she has bought, on average, about eight bottles per month. But she certainly doesn't know when she drank which bottle and so there seems to be no way she can find out, even approximately, the average age of the bottles she has been drinking.

This is a good task for Little's Law.

D. Chhajed and T.J. Lowe (eds.) *Building Intuition*: *Insights From Basic Operations Management Models and Principles*.
doi:10.1007/978-0-387-73699-0, © Springer Science + Business Media, LLC 2008

## *Little’s Law Deals with Queuing Systems*

A “queuing system” consists of discrete objects we shall call “items” that “arrive” at some rate to the “system.” Within the system the items may form one or more queues and eventually receive “service” and exit. Figure 5.1 shows this schematically.

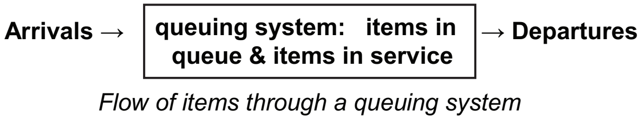

Arrivals →
queuing system: items in queue & items in service
→ Departures
Flow of items through a queuing system

**Fig. 5.1** Schematic view of a queuing system

While items are in the system, they may be in queues or may be in service or some in queue and some in service. The interpretation will depend on the application and the goals of the modeler. For example in the case of the wine cellar, we say that a bottle (an “item”) arrives to the system when it is first placed into the wine cellar. Each bottle remains in the system until Caroline selects it and removes it from the cellar for consumption. If we view the wine rack as a single channel server, the service time is the time between successive removals. It is interesting to note, however, that we do not know which bottle Caroline will pick and there is no particular reason to believe that she will pick according to a first-in, first-out (FIFO) rule. In any case, to deal with the average number of bottles in the cellar or average time spent by a bottle in the cellar, we need to consider the complete system consisting of queue plus service.

Little’s Law says that, under steady state conditions, the average number of items in a queuing system equals the average rate at which items arrive multiplied by the average time that an item spends in the system. Letting

$L$ = average number of items in the queuing system,
$W$ = average waiting time in the system for an item, and
$\lambda$ = average number of items arriving per unit time, the law is

$$L = \lambda W. \tag{1}$$

This relationship is remarkably simple and general. We require stationarity assumptions about the underlying stochastic processes, but it is quite surprising what we do *not* require. We have not mentioned how many servers there are, whether each server has its own queue or a single queue feeds all servers, what the service time distributions are, or what the distribution of inter-arrival times is, or what is the order of service of items, etc.

In good part because of its simplicity and generality, the equation (1) is extremely useful. It is especially handy for “back of the envelope” calculations. The reason is that two of the terms in (1) may be easy to estimate and not the third. Then Little’s Law quickly provides the missing value.

Thus for Caroline, the average number of bottles in the system is $L = (240)*(2/3) = 160$ bottles and the average arrival rate is $\lambda = (12)*(8) = 96$ bottles/year. Without ever collecting individual data on how long each bottle remains in her cellar, she can calculate the average amount of time a bottle stays in her cellar as $W = (160)/(96) \cong 1.67$ years. That's not very old. She needs a bigger rack and more patience, or, alternatively, she should develop selection rules to favor holding special bottles longer than the others. This wouldn't affect the average but might give her some fine old wines.

## Arguing Little's Law with a Picture

Figure 5.2 shows one possible realization of a particular queuing system. We can make a heuristic argument for Little's Law by interpreting the area under the curve in Fig. 5.2 in two different ways. Let

*n(t)* = the number of items in the queuing system at time *t*;
*T* = a long period of time;
*A(T)* = the area under the curve *n(t)* over the time period *T*;
*N(T)* = the number of arrivals in the time period *T*.

On the one hand, an item in the queuing system is simply there. The number of items can be counted at any instant of time *t* to give *n(t)*. Its average value over *T* is the integral of *n(t)* over *T* (i.e., *A(T)*) divided by *T*. On the other hand, at time *t* each of the items is waiting and so is accumulating waiting time. By integrating *n(t)* over the time period *T*, we obtain a cumulative measure of the waiting time, again equal to *A(T)*. Furthermore, the arrivals are countable too, and given by *N(T)*. Therefore, inspecting the figure, we define

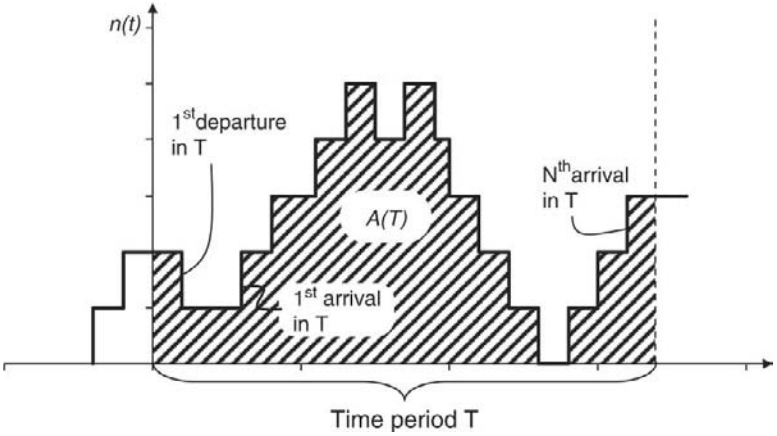


**Fig. 5.2** Number of items in a queuing system versus time

$\lambda(T) = N(T)/T$ = arrival rate during time period $T$,
$L(T) = A(T)/T$ = average queue length during time period $T$,
$W(T) = A(T)/N(T)$ = average waiting time in the system per arrival during $T$.
A slight manipulation gives $L(T) = \lambda(T)W(T)$.

All of these quantities wiggle around a little as $T$ increases because of the stochastic nature of the queuing process and because of end effects. End effects refer to the inclusion in $W(T)$ of some waiting by items which joined the system prior to the start of $T$ and the exclusion of some waiting by items who arrived during $T$ but have not left yet. As $T$ increases, $L(T)$ and $\lambda(T)$ go up and down somewhat as items arrive and later leave.

Under appropriate mathematical assumptions about the stationarity of the underlying stochastic processes, the end effects at the start and finish of $T$ become negligible compared to the main area under the curve. Thus, as $T$ increases, these stochastic "wiggles" in $L(T)$, $\lambda(T)$, and $W(T)$ become smaller and smaller percentages of their eventual values so that $L(T)$, $\lambda(T)$, and $W(T)$ each go to a limit as we increase $T$ to infinity. Then, using the obvious symbols for the limits, we have:

$$\lim_{T\to\infty} L(T) = L; \quad \lim_{T\to\infty} \lambda(T) = \lambda; \quad \lim_{T\to\infty} W(T):$$

from which we get the desired result (1).

It is interesting and important to note that the formula holds for *each* realization of the queuing system over time. This was argued by Little, in his original paper (Little 1961), noting that the relationship (1) held for each evolution of the time series of a particular queuing system. In other words, if we watch a specific case or realization designated, say, by $\omega$, as it develops over time, then we will find that

**Building Intuition** Little's Law provides a fundamental relationship between three key parameters in a queuing (or waiting line/service) system: the average number of items in the system, the average waiting time (or flow time) in the system for an item, and the average arrival rate of items to the system. The system can be very general. For example, it might include both the service facility and the waiting line, or it might be only the waiting line. An important feature of Little's Law is that by knowing, perhaps via direct measurement, two of the three parameters, the third can be calculated. This is an extremely useful property since measurement of all three parameters may be difficult in certain applications.

Little's Law is applicable in many environments including manufacturing and service industries as well as everyday decision making by individuals.

$L(\omega) = \lambda(\omega)\ W(\omega)$, given the steady state and other assumptions made. Averaging cross-sectionally across the many possible realizations of a particular system gives (1), but it is a useful insight to know that the formula holds for each evolving time series as it is observed over a long time period.

## *Law or Tautology?*

Equation (1) is commonly called Little's Law and we have cheerfully adopted that terminology. However, as pointed out by various people, including Little (1992), Eq. (1) is a mathematical theorem and therefore a tautology. The relationship turns out to be useful in practice, but there is no need to go out on the factory floor and collect data to test it. This would be required in the case of a physical law such as Newton's Law of Gravitation. Each side of Newton's equation has to be measured and it is an empirical question whether they are equal within the measurement error. For a mathematical theorem, if the assumptions are satisfied by the application, the result will hold. Note that calling a mathematical theorem a law is not without precedent. The Law of Large Numbers would be another instance.

## Usefulness of Little's Law in Practice

In this section we try to convey the generality of the result and its usefulness in different contexts by means of simple examples. In each case we see how the observation of two of the three measures provides the third. We try to bring out why such back-of-the-envelope analyses are of interest and value in different situations.

***Semiconductor Factory***: Semiconductor devices are manufactured in extremely capital-intensive fabrication facilities. The manufacturing process entails starting with a silicon wafer and then building the electronic circuitry for multiple identical devices through hundreds of process steps. Suppose that the semiconductor factory starts 1,000 wafers per day, on average; this is the input rate. The start rate has remained fairly stable over the past 9 months. We track the amount of work-in-process (*WIP*) inventory. The *WIP* varies between 40,000 and 50,000 wafers; the average *WIP* is 45,000 wafers.

Then we can infer the average flow time in the factory. The arrival rate to the factory is the wafer start rate: $\lambda$ = 1,000 wafers per day. The *WIP* is the system queue length: $L$ = 45,000 wafers. Thus the wait time or expected time in the system is $W$ = 45 days. In a manufacturing context, we often refer to this as the flow time, the time between when a job starts and finishes in a factory. For instance, if we think of one wafer as being a job, then it takes the factory on average 45 days to process it, that is, to convert it from a blank wafer into a finished wafer comprised of electronic devices. Knowing the flow time is critical for planning and scheduling the factory, and for making delivery commitments to customers. We shall return later to the connection between Little's Law and operations management.

***E-Mail***: Managing our e-mail is a common and time-consuming daily activity. For many it is hard to keep up with the volume of messages, let alone provide timely responses. A student Sue might receive 50 messages each day to which she must generate a response. Can we easily assess how well this student handles her e-mail duties?

Indeed we can apply Little's Law to get a quick sense of how promptly Sue responds to messages. Suppose that she receives about 50 messages every day; then this is the arrival rate: $\lambda = 50$ messages/day. Suppose we can also track how many messages have yet to be answered. For instance, suppose that Sue removes a message from her InBox once she has responded to it. Then the remaining messages in her InBox are the messages that are waiting to be answered. Over the last semester, the size of the InBox has varied between one and two hundred messages with an average of 150 messages. Then we can regard this to be the system queue length: $L = 150$ messages. From Little's Law we immediately have an estimate of how long it takes Sue to answer a message, on average: $W = 3$ days.

***Hospital Ward***: We wish to determine the size and staffing levels for the maternity ward for a local hospital. From historical records we know that the birth rate for the local community is about five births per day. We also know that most women stay in the maternity ward for 2 days before going home with child; however occasionally, there are complications with the birth that require much longer stays. Over the past 6 months, we find that 90% of the births have resulted in 2-day stays; for the remaining 10% of the cases, the average length of time in the maternity ward is 7 days. Thus, on average, the length of stay is $0.9 \times 2 + 0.1 \times 7 = 2.5$ days.

We can use Little's Law to predict the average number of mothers in the maternity ward. The arrival process corresponds to the women arriving to deliver their babies; the arrival rate is $\lambda = 5$ mothers per day. The relevant waiting time in the system is the length of stay in the maternity ward: $W = 2.5$ days. Thus, the expected queue length or number in the system is $L = 12.5$ mothers. This would be useful in determining the size of the maternity ward (e.g., beds) and the staffing requirements. However, the law only provides the average requirements, and one would need to design the maternity ward to accommodate its peak requirements. For instance, we would certainly want more than 12.5 or 13 beds in order to handle the variability in the occupancy of the ward. One needs to use queuing models and/or simulation to explore the trade-offs between the utilization of the beds and the likelihood of not having a bed for an expectant mother. Nevertheless, Little's Law provides a starting point for this investigation, since we know the average number of beds that are needed.

***Toll Booths***: The Ted Williams Tunnel travels under the Boston harbor, connecting East Boston to South Boston. During the course of a day, about 50,000 vehicles go through the tolls at the entry point to the Tunnel in East Boston. The Massachusetts

Transit Authority (MTA) tries to modulate the number of toll booths that are open at any point in time so that the average number of vehicles waiting at the tolls (including those at the booths) never exceeds 20 vehicles. For instance, all six booths are open during the peak time in the morning from 6:00 AM to 10:00 AM. During this morning rush hour, the tunnel handles up to 4,000 cars per hour, and the MTA estimates that the average number of vehicles waiting at any point of time is near the target maximum of 20 vehicles.

With the assumption that the arrivals occur at a relatively stable rate over the morning rush hour, we can then use Little’s Law to ask what quality of service is being delivered in terms of average waiting time per vehicle. Suppose that the arrival rate to the toll booths is $\lambda = 3{,}600$ vehicles per hour (or 1 vehicle per second), and the expected number of vehicles in the system is $L = 20$ vehicles. Thus, on average, the time a vehicle spends at the toll booths is $W = 20/3{,}600$ h = 20 s.

***Housing Market***: The local real estate agent in your community estimates that it takes 120 days on average to sell a house; whereas this number changes some with the economy and season, it has been fairly stable over the past decade. You observe from monitoring the classified ads that over the past year the number of houses for sale has ranged from 20 to 30 at any point in time, with an average of 25. What can we say about the number of transactions in the past year?

From Little’s Law we can estimate this by viewing the real estate market as a queuing system. We regard a house being put up for sale as an arrival to the system. We assume that an unsold house remains on the market until it is sold. Thus, when a house “completes its service” and departs from the market, we infer that it has been sold. We have estimates of the average time in the system and the average number in the system, namely, $W = 120$ days and $L = 25$ houses. From this, we can estimate the arrival rate to the system, $\lambda = 25/120$ houses per day $\cong 75$ houses per year.

***Doughnut Shop***: From your daily morning trip to the doughnut shop, you know they have a healthy business, at least financially speaking. As you might want to invest in a franchise, you wonder what amount of revenue they generate. Over the course of several months, you visit the shop at random times between 6:00 AM and 9:00 AM; you observe that the queue averages about 10 customers, and that it takes you about 3 min to get in and out of the shop.

If you assume your experience is typical, then you can apply Little’s Law to estimate what the throughput rate is for the enterprise for the morning peak period. The expected number in the system is $L = 10$ customers and the expected time in the system is $W = 3$ min. We can then estimate the arrival rate to the system, namely $\lambda = 10/3$ customers per minute = 200 customers per hour. We also term this the throughput rate as arriving customers become throughput once served. To get an estimate for the revenue potential from this shop, we need to estimate how much each customer spends. If you typically spend $5 per visit, then with the assumption that you are a typical customer, we have a rough estimate of the shop’s revenue during these morning hours, i.e., $1,000 per hour.

## The Robustness and Generality of Little's Law in Certain Systems

So far we have developed and discussed Little's Law as a relationship among steady-state stochastic processes. The contexts we have examined have been well-behaved, stable, and on-going. In particular we assume that the characteristics of the arrival and service processes are stationary over time. For example, in the case of the maternity ward, we assume that the average arrival rate of mothers has been steady at five per day for some time, and that this rate does not vary with day of week or season of the year. Similarly, we have regarded the service process as being stationary; for instance, we read and process our e-mail at roughly the same average rate, day in and day out, independent of the backlog of unread messages. For some of our examples, we have focused on an interval of time, e.g., the morning rush hour through the toll booths. However, in these instances due to the huge volume of arrivals, we contend that the system behavior is virtually equivalent to that of a steady-state system.

The purpose of this section is to show the great robustness and generality of Little's Law under certain circumstances. Indeed Little's Law is exact in these cases even though arrival and service process may be nonstationary. The essential condition is to have a finite window of observation that starts and stops when the system is empty. We use an example to motivate and illustrate the validity of Little's Law in this situation. Consider the *Sweet & Sour* supermarket, which opens every day at 7:00 AM and closes 16 h later at 11:00 PM. When *S&S* opens at 7 AM, there are no customers in the store. When it closes at 11 PM, all of the customers depart. Between opening and closing, customers arrive to the store, do their shopping and leave. The arrivals over the course of the day are quite varied. They include several customer segments, each with quite distinct shopping habits. Families with school-age children will shop between 9AM and 2 PM, and tend to have fairly lengthy shopping forays as they stock up for a week at a time. Seniors will tend to shop at quiet times of the day, like first thing in the morning, and will also be fairly leisurely in their shopping, taking up to an hour to complete a visit. Working couples will shop at night after work or on the weekends; their evening visits are often to run in, grab something and run out.

We propose to model *S&S* as a queuing system with the arrivals being the customers as they enter the store and service being the duration of their time in the store selecting and purchasing their groceries. However, from the above discussion, we see that this is anything but a stable system. The supermarket is never in a steady-state. It starts and ends each day with zero customers. Over the course of the day, customers arrive at varying rates, and the nature of their shopping trips also varies over the day, due to the different clienteles. Nevertheless, we will show next that Little's Law applies each and every day to this supermarket in an exact way.

## *An Analytic Interlude*

Let $N$ denote the number of customers that shop on a particular day. Suppose that we keep track of when customers arrive and when they depart from the store. Then we can define and create two processes, one for the arrivals and the other for the departures. We define time $t = 0$ to correspond to the opening at 7 AM, and time $t = 16$ to be the store closing at 11 PM, 16 h later.

We let $N(t)$ denote the cumulative number of arrivals to the store by time $t$. Thus, as we start the day with zero customers, we have $N(0) = 0$; as we assume a total of $N$ customers arrive during the day, we have $N(16) = N$. The cumulative arrivals increase in a stair-case fashion, as shown in Fig. 5.3, over $0 < t < 16$.

In similar manner we define $D(t)$ to denote the cumulative number of departures from the store by time t. Again, we have $D(0) = 0$, $D(16) = N$, and the cumulative departures increase in a stair-case fashion over the time interval $0 < t < 16$; see Fig. 5.3.

We note that at all time instants we have $N(t) \geq D(t)$, as the number of departures can never exceed the number of arrivals. Indeed, the difference between the two cumulative processes is the number of customers in the supermarket at time t:

$$L(t) = N(t) - D(t).$$

With this observation we can determine the average number in the supermarket over the course of the day from the following integral:

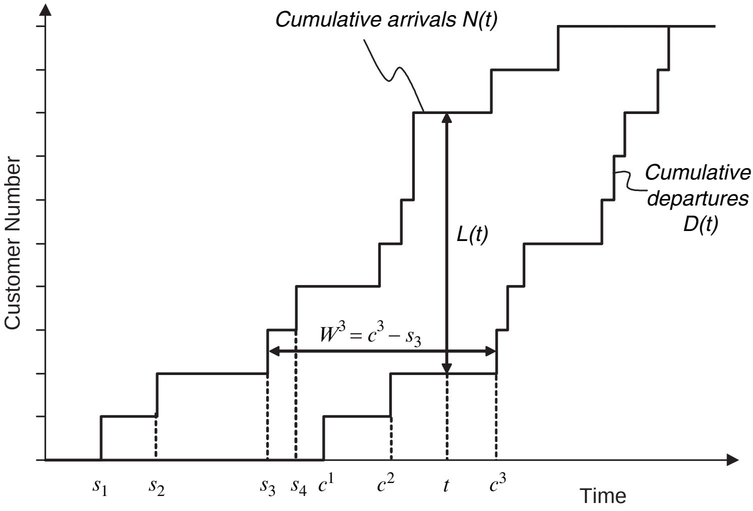


**Fig. 5.3** Cumulative arrivals to and departures from a system, for example, the supermarket

$$L = \frac{1}{16} \times \int_{t=0}^{t=16} (N(t) - D(t))dt. \tag{2}$$

To model the average time in the supermarket for each customer is a bit more involved. We define $\{s_1 ,s_2 , \ldots s_N \}$ to be the sequence of arrival or start times for the $N$ customers, where $s_j$ denotes the start time for the jth arriving customer. We define $\{c_1 ,c_2 , \ldots c_N\}$ to be the sequence of departure or completion times for the $N$ customers, where $c_j$ denotes the completion time for the jth arriving customer. Thus, the time in the supermarket for the jth arriving customer is $W_j = c_j - s_j$. Averaging this over all the customers gives:

$$W = \frac{1}{N} \times \left( \sum_{j=1}^{N} c_j - \sum_{j=1}^{N} s_j \right). \tag{3}$$

To compute (3) we shall develop an equivalent expression based on the geometry in Fig. 5.3. Let us define $\{c^1,c^2 ,\ldots c^{\mathrm{N}}\}$ to be the sequence of departure or completion times for the $N$ customers, where $c^j$ denotes the completion time for the jth departing customer. Since customers need not exit the store in the order that they arrive, we shall often have $c_j \neq c^{\mathrm{j}}$. However, the sequence $\{c^1, c^2 ,\ldots, c^{\mathrm{N}}\}$ is just a permutation or reordering of the sequence $\{c_1 ,c_2 ,\ldots c_{\mathrm{N}}\}$ as the departure time for each customer must appear exactly once in each sequence.

As shown in Fig. 5.3, we can define the jth wait time as $W^{\mathrm{j}} = c^{\mathrm{j}} - s_j$, equal to the difference between the departure time for the jth departing customer and the start time for the *j*th arriving customer. Now let us consider the average of these wait times:

$$\frac{1}{N} \times \sum_{j=1}^{N} W^j = \frac{1}{N} \times \left( \sum_{j=1}^{N} c^j - \sum_{j=1}^{N} s_j \right).$$

But this will equal $W$ given by (3), since $\sum_{j=1}^{N} c_j = \sum_{j=1}^{N} c^{\mathrm{j}}$. Hence we conclude that

$$W = \frac{1}{N} \times \sum_{j=1}^{N} W^j = \frac{1}{N} \times \left( \sum_{j=1}^{N} c^j - \sum_{j=1}^{N} s_j \right). \tag{4}$$

Now we need to relate our expression for $L$, given by (2) to our expression for $W$, given by (4). From the geometry in Fig. 5.3, we observe the following equivalence:

$$\int_{t=0}^{t=16} (N(t) - D(t))dt = \sum_{j=1}^{N} W^j. \tag{5}$$

That is, on each side of the equation, we have an expression for the area between the cumulative arrivals and the cumulative departures. On the left side we compute the area by integration over time of the function that tracks the number in the system; on the right side we compute the area by summing up the time in system for *N* customers.

From (2), (4) and (5), we can now write Little's Law for the supermarket:

$$L = \frac{1}{16} \times \int_{t=0}^{t=16} \left(N(t) - D(t)\right) dt = \frac{1}{16} \times \sum_{j=1}^{N} W^{j} = \frac{N}{16} \times W. \tag{6}$$

We recognize $\frac{N}{16}$ to be the arrival rate in customers per hour for the particular day, and we define $\lambda = \frac{N}{16}$; thus we have (6) in its familiar form, $L = \lambda W$.

With this simple example we have shown that Little's Law can be true over a finite time window (16 h) with nonstationary arrivals and with no notion of any steady state for the system in question. On reflection, there were two essential conditions for this result:

- Boundary conditions—we specify the finite time window to start and end with an empty system. This was a natural condition for the supermarket, and indeed, would be common for many service systems.
- Conservation of customers—we assume that all arriving customers will eventually complete service and exit from the system; there are no lost customers, so that the number of arrivals equals the number of departures. Again, this is a valid assumption for many systems of interest.

We really needed nothing else beyond these conditions in order to establish the law in our case. We have no assumptions about *N*, the number of customers; indeed, all the equations hold true for any *N*, e.g., for *N* = 1. We have no assumptions about the process for arrivals, or about how customers are serviced within the store. There might be a long period of no arrivals followed by the arrival of several busloads of customers; there might be periods of no service completions, say, if all the cash registers stopped working for an hour. The only conditions are as stated above: we need to start and end with an empty system and we need to conserve customers.

Notice that our formula is exact, but after the fact. In other words, we cannot complete our calculation until the supermarket door shuts. This is not a complaint. It merely says that we are observing and measuring not forecasting. Another point to mention is that the numbers will be different each day because of different sets of shoppers on different days of the week, the weather, holidays, and other changes in the store's internal and external environment. Nevertheless, the relationship $L = \lambda W$, as measured for that day, will be exact and the ability to measure two of the parameters and deduce the third still holds.

## *Further Discussion of "Average"*

Little's Law holds exactly, but let us examine further what we mean by "average" wait, queue, and arrival rate. We have no probability distributions and so these are not expected values. Looking at the derivation of the result, we see that we are talking about everyday sample averages in the case of waits and arrival rates and finite time averages in the case of queues. So Little's Law here shows us an exact relation among *sample and time averages*. Next, consider a customer segment at the *Sweet & Sour* supermarket that consists of men with children in strollers. We can compute sample or time averages for their arrival rate, time in store, and number in store. Little's Law will hold exactly. Therefore, Little's Law is true for these averages for any identifiable segment. To use the relationship in practice, it will be necessary to collect data observing how many people of the target segment enter the store during the day.

To summarize, Little's Law is robust and remarkably general for queuing systems for which a finite window of observation starts and stops when the system is empty. Interpretation of the area between cumulative arrivals and cumulative departures permits an analytic argument that Little's Law is exact despite possibly nonstationary arrival and service processes. What we have discussed here turns out to be the tip of a fascinating mathematical iceberg that has been developed in recent years, called sample path analysis of queuing systems. An adequate discussion of it is beyond the scope of this chapter but the interested reader is referred to the book of El-Taha and Stidham (1999).

## Evolution of Little's Law in Operations Management

Over the past 15 years or so, Little's Law has played an increasingly important role in the teaching and practice of operations management. However, the law is usually stated in a modified format to emphasize its applicability to operations. For instance, we cite as an example the very successful textbook of Hopp and Spearman (2000) who refer to Little's Law as a "…an interesting and fundamental, relationship between *WIP*, cycle time and throughput." They go on to state the law as

$$TH = \frac{WIP}{CT}, \tag{7}$$

where they define throughput (*TH*) as "the average output of a production process (machine, workstation, line, plant) per unit time," work in process (*WIP*) as "the inventory between the start and end points of a product routing," and cycle time (*CT*) as "the average time from release of a job at the beginning of the routing until it reaches an inventory point at the end of the routing (that is, the time the part spends as *WIP*)." They note that cycle time is also referred to as flow time, throughput time, and sojourn time, depending on the context.

We easily see that (7) is equivalent to Little’s Law with $TH = \lambda$, $WIP = L$ and $CT = W$. However, there is a more fundamental difference in that the law is stated in terms of the average output or departure rate for the system, rather than the arrival rate. This reflects the perspective of a typical operating system, especially a manufacturing operation. Output is a primary attribute of any manufacturing system, since it is nominally its *raison d’etre*. As stated, we see that any increase in output requires either an increase in work-in-process inventory or a reduction in cycle time or both.

Furthermore in many contexts, the output rate is determined exogenously and is given to the manufacturing system; it reflects actual sales and/or a forecast of sales. The manufacturing system must then manage its operations to achieve this output rate. It will need to determine how to release work to the operation so as to meet the output target. In effect, the arrival process is endogenous. The operations manager decides the arrivals to the system based on the desired outputs. There is extensive research in the operations literature on how best to set the work release (or arrival process) to achieve the output targets. The best policies are dynamic policies that depend on the state of the manufacturing shop, e.g., depend on the work-in-process.

Our original development of Little’s Law assumes a stable system with a stationary arrival process; as discussed above, we cannot assume a stationary arrival process for the typical context in which we might apply (7). Thus, we ask what conditions are necessary for (7) to be valid. At a minimum we must have conservation of flow. Thus, the average output or departure rate (*TH*) equals the average input or arrival rate ($\lambda$). Furthermore, we need to assume that all jobs that enter the shop will eventually be completed and will exit the shop; there are no jobs that get lost or never depart from the shop. In addition, we need some notion of system stability. We consider two possibilities, as this issue raises another important consideration.

First, we might assume that the shop will occasionally empty, i.e., $WIP = 0$. Then, as with the supermarket example, we will see that Little’s Law holds exactly between any two time instances at which the shop is empty.

However, in many manufacturing systems, the *WIP* never drops to zero. In some contexts, this occurs for behavioral reasons; as the *WIP* decreases, the shop naturally slows down so as to not run out of work. The shop adjusts its service rate dynamically so as to keep from driving the *WIP* to zero. In other contexts, there might be an explicit control rule that maintains some target level or range of *WIP*. For instance, a very effective control policy is the so-called CONWIP policy that maintains an absolutely constant level of *WIP* (Hopp and Spearman, 2000); that is, the control rule releases one unit of new work to the system whenever one unit of work completes processing and exits the system. In either case, Little’s Law applies, at least as an approximation, as long as we select a time interval that is long enough for two conditions to hold.

First, we need the size of the *WIP* to be roughly the same at the beginning and end of the time interval so that there is neither significant growth nor decline in the size of the *WIP*.

Second, we need some assurance that the average age or latency of the *WIP* is neither growing nor declining. We have previously assumed that all jobs that enter the shop will eventually be completed and will exit the shop. But if the *WIP* never drops to zero, it is possible for the jobs to be getting older or younger, in which case the law does not hold.

To illustrate the problem, consider a doll store that offers a line of international folk dolls. Each doll is adorned with traditional folk clothes from a particular country. The artist who produces the dolls gives each doll a distinctive hat in one of 10 different colors, which span the rainbow; the choice of color is completely at random. So, sometimes the Irish folk doll has a green hat, and sometimes a red hat. The store stocks 100 of these dolls so as to have a rich assortment of dolls from which to choose; whenever it sells a doll, the store immediately obtains a replacement from the supplier so as to maintain its in-store stock at 100 dolls. The supplier chooses the replacement doll at random from its supply.

Demand for the dolls has been quite good, as customers appreciate the artistry and novelty of the dolls. Indeed, the dolls sell consistently at a rate of about two per week. However, customers have a subtle but strong subconscious dislike for dolls with mauve hats. Dolls with mauve hats seldom sell; indeed, these mauve-hat dolls sell at a rate of about one every 2 years.

A naïve application of Little's Law might assume a doll arrival rate, equal to the demand rate, $\lambda = 2$ per week and an average number in system $L = 100$ dolls, and then conclude that the average time in the store for a doll is $W = 50$ weeks. However, this is not likely to be an accurate estimate over any moderate time interval, like a few years. Suppose we start with none of the hundred dolls having mauve hats. Every time we sell a doll there is a 1 in 10 chance that it will be replaced with a doll with a mauve hat, as the artist has 10 colors from which to choose. Since these mauve-hat dolls sell much, much less frequently than any other doll, the mix of dolls will gradually change over time. Indeed, the number of mauve-hat dolls in the store grows by about 10 dolls per year. Dolls without mauve hats make up a smaller percentage, but continue to sell at a rate of two per week; the time in system for these dolls actually shrinks as they make up a smaller percentage of the store assortment. However, the dolls with mauve hats do not sell and just accumulate more and more waiting time. Hence, the average age of the dolls in the store's assortment continues to grow older until at some point, the entire assortment has mauve hats. As a consequence, we cannot apply Little's Law during this transient period.

For instance, suppose we observe the system for 5 years, and then use Little's Law to estimate $W = L/\lambda = 100/2 = 50$ weeks; this estimate overstates the actual time in system for those dolls that have sold. Over the first 5 years, the number of mauve-hat dolls in the store grows from zero to about 50. As a consequence the active inventory, namely the dolls without mauve hats, drops from 100 to about 50. These dolls stay in the system, on average, less than 50 weeks, whereas the mauve-hat dolls just sit and get older. Of course, if we extend the time interval, in 10 years the entire assortment becomes mauve and the demand rate falls to one doll every 2 years; eventually (about 200 years) the mauve inventory turns over and Little's Law will now apply. Presumably, the store would recognize the trend a bit earlier and do something about it! Nevertheless, the example shows how Little's Law might not hold over some time interval during which the average age of the *WIP* or queue is changing.

A quite different approach to the problem of nonzero *WIP* is motivated by what we learned in the supermarket problem. There we noted that Little's Law applies independently to each customer segment, but we have to be able to identify the customers in each segment and collect data on them.

So, in the case of non-zero *WIP* (or, actually, any existing *WIP*!), we can ignore all of it and focus on a group of new items, which, for example, might be colored blue. The system starts empty of blue items, even though it may be cluttered with others. We count blues as they enter the system (the rate may be controlled if we wish—stationarity is not required). The observations we make depend on what we want to learn and what is easy to measure. Suppose we want to observe and process $N$ blue items. And suppose we want to know the cycle time $CT = W$, the average time a blue item spends in processing. At regular intervals we take an inventory of blues so that we can estimate $WIP = L$ (for blues only) by simple averaging. Eventually all $N$ leave, say at $T$ and so the average arrival rate is $\lambda = TH = N/T$. Then $CT = W$ can be calculated by Little's Law. The underlying theory is exact, although we have introduced some sampling error in estimating the blue *WIP*. However, this is something we can control by putting whatever resources we think are worthwhile to reduce it.

Alternatively, we might have a way of determining blue cycle time $CT = W$ exactly and wish to know the average inventory of blue items, i.e., the blue $WIP = L$. This is the case in the following example.

Consider a toy manufacturer that contracts with a third party logistics firm, 3PL, to handle its on-line business. The toy manufacturer will supply inventory to 3PL to fill orders. When the toy manufacturer receives an order on-line, it will instruct 3PL to fill it.

The toy manufacturer pays 3PL to provide this service. The contract terms depend on two factors: the number of orders shipped and the amount of inventory space occupied by the toys at 3PL. The toy manufacturer pays 3PL $10 for every order that is shipped and $0.03 per day for each unit in inventory.

At the start of each month, the toy manufacturer ships a batch of toys to 3PL. These toys typically sell out within the month, but not always. For accounting reasons, the toy manufacturer insists that 3PL submit an invoice for its services for each batch of toys. Hence, 3PL waits until the entire batch has been sold before it can finalize an invoice. In preparing the invoice 3PL can easily determine the shipping-cost component, as it just depends on the number of toys in the batch. However, 3PL is less clear about how to account for how much inventory space is attributable to a batch of toys. One approach would be to count its inventory each day. However, it would be quite difficult to track the inventory associated with a particular batch since the toys from all batches, as well as from other suppliers, are stored together in one section of the warehouse. Thus, it would be cost prohibitive to do an inventory count each day.

A much simpler approach is to use Little’s Law. Suppose that whenever 3PL receives a batch of toys, it records the arrival date for each toy. 3PL can also record the date at which the toy is shipped, and hence obtain the time in system for each toy.

For instance, if there were an RFID tag attached to each toy, then 3PL can easily record these transactions with RFID readers at its shipping docks. Thus, 3PL knows the wait times $W_i$ for $i = 1,2, \ldots N$, where $W_i$ is the difference between the ship time and receipt time for the $i^{\text{th}}$ toy and $N$ is the number of toys in a batch. The average wait time for the batch is

$$W = \frac{1}{N}\sum_{i=1}^{N} W_i.$$

If it takes $T$ days to sell the batch of $N$ toys, then the average arrival rate for the batch is

$$\lambda = \frac{N}{T}.$$

By Little’s Law we now can find the average inventory attributable to this batch:

$$L = \lambda W = \frac{1}{T}\sum_{i=1}^{N} W_i.$$

Since $L$ represents the average inventory over a period of $T$ days, 3PL will charge the toy manufacturer for L × T unit-days of inventory storage, at $0.03 per unit-day.

## Concluding Remarks

We have given a variety of examples showing the kinds of situations where Little’s Law can usefully convert an estimate of an average queue into an estimate of average waiting time and vice versa when one may be relatively easy to measure and the other not.

We have also briefly examined how Little’s Law has been used in operations management. Here we observe that a different terminology and set of symbols is usually adopted. Operations managers are concerned with their throughput rate rather than an arrival rate; their queue length is usually *WIP*, their wait time is termed cycle or flow time. We also discussed two differences in orientations between the use of the law in operations management and its original derivation and application:

- For operations management the law is often expressed in terms of output rates rather than arrival rates. Furthermore, the arrival process to most operating systems is not a stationary process, and, indeed, may be controlled.
- Many operating systems are never empty. That is, the number in the system or *WIP* is always positive.

In each case we see that Little's Law can apply, albeit with some required conditions and thoughtful attention to the goals of the application.

## Historical Background

We trace the evolution of Little's Law from the early days of queuing theory in operations research and management science to our own chapter in this book.

The earliest paper we have found that makes use of Little's Law simply assumes it to be true. Cobham (1954), in an article on priority queues, writes, "… it is sufficient to observe that the *expected number* of units *of priority k* waiting to be serviced is $\lambda_k W_k$, where $W_k$ *is the expected wait* for a unit of priority *k*." ($\lambda_k$ is the arrival rate of priority *k* items.)

We attribute the explicit formula "$L = \lambda W$" to Philip Morse and his book, *Queues, Inventories and Maintenance*, (Morse, 1958). He does not give the law a name but simply talks about "the relation between mean number and mean delay." In Chap. 7 he proves the relationship for a single channel queuing system having Poisson arrivals and a service time distribution of the general class known as k-Erlang. The proof of $L = \lambda W$ is, in a certain sense, incidental to his main task of solving for the steady-state joint probabilities of the number of items in the system and the stage of item in service. Using the fairly standard approach of describing the system with a set of differential equations, Morse solves the system by building a two variable generating function of the desired joint probabilities. One variable relates to the number of items in the system and the other to the stage of the item in service. The generating function can then be used rather easily to find both *L* and *W*. Examination shows that *L* and *W* differ only by the constant of proportionality λ. This is a nice piece of analysis of the particular system, but it does not readily suggest a route to greater generality.

Morse was clearly interested in the general case. After the above analysis he wrote, "we have now shown that... the relation between the mean number and mean delay is via the factor λ, the arrival rate: $L = \lambda W$ and..... We will find, in *all* the examples encountered in this chapter and the next, for a wide variety of service and arrival distributions, for one or for several channels, that this same relationship holds. Those readers who would like to experience for themselves the slipperiness of fundamental concepts in this field and the intractability of really general theorems, might try their hand at showing under what circumstances this simple relationship between *L* and *W* does *not* hold."

The next paper to appear was Little (1961) and had the title “A Proof of the Queuing Formula: $L = \lambda W$.” Little essentially analyzes the picture shown in Fig. 5.2 of our chapter. His argument uses ergodic theorems for strictly stationary stochastic processes, as drawn from Doob (1953). The basic analytic task is to get rid of “end effects” as discussed below Fig. 5.2 in this chapter. However, Little does something else not done previously. His proof argues that the law holds in any specific realization of the queuing system when observed over a long period of time. This is different from finding steady state probabilities for the number in the system and calculating $L$ and performing a corresponding analysis of the distribution of waiting times to find $W$. The idea that the theorem is true for each evolution of the system provides a deeper understanding of the importance of the relationship $L = \lambda W$ in the queuing process itself. It also lays a basis for sample path proofs of the relationship that are to come.

Not too long afterward, Jewell (1967) came along with “A Simple Proof of $L = \lambda W$.” He draws a picture very similar to our Fig. 5.3 and makes the assumption that the event when the system becomes empty is a recurrent event. By definition the times between such events are mutually independent random variables having the same distribution. Thus the time line alternates between intervals during which the system is empty and ones during which the system is busy. Jewell assumes that arrival and waiting mechanisms are reset at the start of each busy period. In other words, in each busy period, the random variables in those mechanisms have the same joint distribution, and their values are new random draws, independent of previous busy periods. An advantage of Jewell’s paper was that it used a vocabulary more familiar to queuing system analysts than the measure-theoretic arguments of stationary stochastic processes invoked by Little (1961).

Jewell’s paper was followed by Eilon (1969), which had the title “A Simpler Proof of $L = \lambda W$.” He shows essentially the same picture as Jewell and as our Fig. 5.3 and makes the same heuristic analysis that we do under Fig. 5.2. He notes that, if the limits of $L(T)$, $W(T)$, and $\lambda(T)$ exist as $T$ goes to infinity, the result is proven. Most writers on the subject, however, consider that a further argument is required to be certain that end-effects go away in the limit.

As time went on, the number and importance of applications of queuing systems increased and with them the applications of Little’s Law. One major area lay in the design and analysis of computer systems. Kleinrock (1975, 1976) develops and summarizes a set of tools to assist this. One of the present authors (Little) recalls receiving a telephone call out of the blue from a computer engineer on the West Coast during the 1970s. The caller asked, “Do you have any more laws? I use Little’s Law all the time and find it really helpful.” But the response from the East Coast was “Sorry, we’re fresh out of new laws today.”

Although researchers in the field had no doubt about the remarkable generality of Little’s Law, there developed among some of them the belief that its validity could be proven deterministically by sample path analysis. Such an approach would produce a reader-friendly proof without recourse to probability, somewhat in the manner that this chapter argues the deterministic validity of Little’s Law in the supermarket example. The net result was Stidham (1974), “A Last Word on $L = \lambda W$.”

However, as he pointed out in a later review article (Stidham, 2002), it was not the last word at all because the research potential of sample path analysis for queuing and other applications has ramifications far beyond Little's Law. Many results of this sort appear in a book by Stidham and a colleague: (El-Taha and Stidham, 1999): *Sample-Path Analysis of Queueing Systems*.

Over the past few years, Little's Law has become increasingly important in operations management. The notation and type of thinking are different, but applications are growing in number and importance. The goal of this chapter has been to facilitate such applications by providing illustrative examples, especially ones that explore new territory.

## *A Note of Personal History (Little)*

How did a sensible young PhD like me get involved in a crazy field like this? From 1957–1962, I taught operations research at the Case Institute of Technology in Cleveland (now Case Western Reserve University). I was asked to teach a course on queuing. OK. Initially I used my own notes, but when Morse (1958) came out, I used his book extensively. Queuing was taken by most of the OR graduate students and, indeed, one of these, Ron Wolff, went on to become a first class queuing theorist (Wolff 1989). One year we were at the point when we had done the basic Poisson-exponential queue and moved through multi-server queues, and some other general cases. I remarked, as many before and after me probably have (and Morse does), that the often reappearing formula $L = \lambda W$ seemed very general. In addition I gave the heuristic proof that is essentially Fig. 5.2 at the beginning of this chapter. After class I was talking to a number of students and one of them (Sid Hess) asked, “How hard would it be to prove it in general?” On the spur of the moment, I obligingly said, “I guess it shouldn't be too hard.” Famous last words. Sid replied, “Then you should do it!”

The remark stuck in my mind and I started to think about the question from time to time. Clearly there was something fundamental going on, since, when you draw the picture you do not really seem to need any detailed assumptions about interarrival times, service times, number of servers, order of service, and all the other ingredients that go into the panoply of queuing models. You only seemed to need a process that goes up and down in unit amounts and some guarantee of steady state and conservation of items. In addition, because I could see there were end effects in the picture, there needed to be a way to get rid of them in the limit. It seemed to me I was in the general arena of stationary stochastic processes. I am not a mathematician by training, and so I bought copies of measure theoretic stochastic process books like Doob (1953), which mentioned stationary processes and ergodic theorems.

My family's habit at the time was to go to Nantucket in the summer where my wife's family had a small summer house. We would load our children in a station wagon, drive to Woods Hole, take the ferry, and spend a couple months away from the world. Since the beach was the baby-sitter, I was able to split off solid blocks of time to work on research as a good assistant professor should. (I wish I could do that today!) I always brought a pile of books and projects with me. $L = \lambda W$ was one of them. I soon ran into problems that required more than looking up theorems in my new books, but I worked out approaches to the road blocks and eventually wrote everything up, giving it my best shot. I sent the paper off to *Operations Research.* It was accepted on the first round.

Nevertheless, I had learned my lesson. I decided that Borel fields and metric transitivity were not going to be my career and retired from queuing. That was in 1961. My retirement held until 2004 when I was accosted by email and in person at an INFORMS meeting by Tim Lowe. Even then, as he will tell you, I resisted re-entrapment, saying, “I don’t know anything about OM and I haven’t looked at $L = \lambda W$ for 40 years.” Being always susceptible to a new challenge and, more importantly, thanks to much help from Steve Graves, who really does know OM, I took a run at holding up my end of the chapter. It has been a wonderful experience and I have learned much. Now I need to find ways to use it.

## References

Cobham, A. (1954) “Priority Assignment in Waiting Line Problems,” *Operations Research*, 2, (1) (Feb.), 70–76.

Doob, J. L. (1953) *Stochastic Processes*, John Wiley, New York.

Eilon, S. (1969) “A Simpler Proof of L = λW,” *Operations Research*, 17, (5) 915–917.

El-Taha, M. and S. Stidham (1999) *Sample-Path Analysis of Queueing Systems,* Kluwer Academic, Boston MA.

Hopp, W. J. and M. L. Spearman (2000) *Factory Physics: Foundations of Manufacturing Management*, 2nd (ed.), Irwin/McGraw Hill, New York, NY.

Jewell, W. S. (1967), “A Simple Proof of L = λW,” *Operations Research*, 15, (6) 1109–1116.

Kleinrock, L. (1975) *Queueing Systems, Volume I: Theory*, A Wiley-Interscience Publication, New York.

Kleinrock, L. (1976) *Queueing Systems, Volume II: Computer Applications*, A Wiley-Interscience Publication, New York.

Little, J. D. C. (1961) “A Proof of the Queuing Formula: L = λW,” *Operations Research*, 9, (3) 383–387.

Little, J. D. C. (1992) “Are there ‘Laws’ of Manufacturing,” *Manufacturing Systems: Foundations of World-Class Practice*, edited by J. A. Heim and W. D. Compton, National Academy Press, Washington, D.C., 180–188.

Morse, P. M. (1958) *Queues, Inventories and Maintenance,* Publications in Operations Research No. 1, John Wiley, New York.

Stidham, S., Jr. (1974) “A Last Word on L = λW,” *Operations Research*, 22, (2) 417–421.

Stidham, S., Jr. (2002) “Analysis, Design, and Control of Queueing Systems,” *Operations Research*, 50, (1) 197–216.

Wolff, R. W. (1989) *Stochastic Modeling and the Theory of Queues*, Prentice-Hall, Inc., Englewood Cliffs, New Jersey.

# Chapter 6
# The Median Principle

**Timothy J. Lowe* and Dilip Chhajed****
***University of Iowa**
****University of Illinois**

*The mean or average value is often used as the representation of a list of numbers when in fact another measure, the median, which is the middle value of the list, is more appropriate. This principle of "middle" is more general and has several other applications—from location to inventory problems.*

## Introduction

Professor Lewis teaches operations management classes at Central College and is known for giving challenging assignments to his students. Because of the amount of work required in completing homework exercises, he has asked students to work in teams and to submit a single paper for grading. Marc, Tej, and Carrie, students in Professor Lewis's class, are in the same team and have decided to meet every Monday and Wednesday to work on Professor Lewis's assignments. Each of them will take a "first stab" at the assigned problems, and then they will meet to compare answers and to finalize their joint paper. Marc lives at 101 Main Street while Tej lives three blocks east of Marc at 401 Main Street. Carrie lives 17 blocks east of Tej at 2101 Main Street. An immediate issue is to choose a location to meet to compare their answers to the problems. The students realize that some walking will be involved, but they are looking for an "efficient" solution to the meeting place problem.

## The Location Problem

So where should the students meet? First, they should decide on an objective to use in making this decision. One possibility is to find a location where the student that walks the farthest would have the least possible walking distance. They observe that Marc and Carrie live 20 blocks apart and Tej lives in between these two students (see Fig. 6.1). Thus if they meet at the midpoint between Marc and Carrie (at 1101 Main Street), the student that walks the farthest will walk 10 blocks. In fact, at this

D. Chhajed and T.J. Lowe (eds.) *Building Intuition*: *Insights From Basic Operations Management Models and Principles*.
doi:10.1007/978-0-387-73699-0, © Springer Science + Business Media, LLC 2008

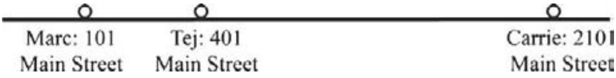


**Fig. 6.1** Location of houses of Marc, Tej, and Carrie

"center" solution to the location problem both Marc and Carrie will walk 10 blocks while Tej will walk 7 blocks. This seems like a good *equitable* solution but is this how they should think about the problem? Is this equitable solution requiring an excessive amount of walking? That is, is this solution efficient?

Suppose an efficient solution is defined by the one which makes the average walking distance as low as possible. The average might be more appropriate in the case where the "meeting problem" is repeated a number of times, e.g., multiple assignments over the semester. Note that for the meeting point at 1101 Main Street, the *average* distance walked by a student is (10 + 10 + 7)/3 = 9 blocks. Thinking that there may be a better solution (to reduce the average walking distance) they begin to do some experimentation. If they decide to meet at 1001 Main Street (one block closer to both Marc and Tej) instead of 1101, Carrie, unfortunately, will walk one additional block but Marc and Tej will each walk one less block. Thus the average walking distance would be 9 + (1−2)/3 = 8 2/3. In fact, note that for each block of movement to the west of Carrie's house, until Tej's house is reached, the average walking distance decreases by 1/3 block. Once the potential meeting location is at Tej's house, any further movement to the west increases Tej's and Carrie's walking distances while Marc's decreases (as long as the location remains east of his house). Thus it seems that the best location is Tej's house. The average walking distance is then (3 + 0 + 17)/3 = 6 2/3 blocks.

Now suppose we consider not only walking distance, but also walking effort as measured by the load each student must carry. Let us suppose that in addition to individual solutions to the homework problems, Marc will carry liquid refreshment for them to share (a cooler weighing 3 pounds). Tej will carry an ultramodern laptop computer weighing 2 pounds. Carrie has a set of books along with some antiquated technical reports that the group may need, giving a total weight of 6 pounds. The total effort a student must expend in reaching the meeting place is related to the weight carried times the distance traveled. Thus if Tej walks nine blocks, his effort is 9 blocks × 2 pounds = 18 lb-blocks and if Carrie walks two blocks, her effort is 2 blocks × 6 pounds = 12 lb-blocks. With this objective in mind the three student may wish to find a location that minimizes the average weighted distance each student must travel. Considering Tej's house as the location, the total weighted distance would be 3 blocks × 3 pounds (for Marc) + 0 blocks × 2 pounds (for Tej) + 17 blocks × 6 pounds (for Carrie). Thus the average would be (9 + 102)/3 = 37. Would this be (as in the first scenario) the location minimizing the average? Considering first a movement one block west of Tej's house, we see that Tej and Carrie's weighted distance contributions increase by $1 \times 2$ pounds and $1 \times 6$ pounds, respectively, while Marc's would decrease by $1 \times 3$ pounds. Thus the average would

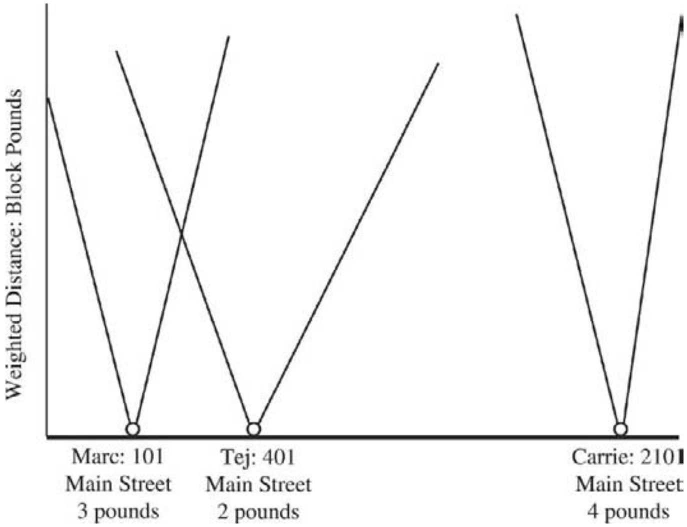


**Fig. 6.2** The weighted distance traveled from house

actually increase by [(1×2) + (1 × 6)−(1 × 3)]/3 = 5/3 block-pounds. However, considering a location one block east of Tej's house (toward Carrie's house) would *decrease* the average weighted distance by [1 × 6 pounds (for Carrie) − 1 × 3 pounds (for Marc) + 1 × 2 pounds (for Tej)]/3 = 1/3 block-pound. In fact, continued movement east results in decreased average weighted distance until Carrie's house is reached. Locating at Carrie's house the average weighted travel distance is [20 × 3 pounds + 17 × 2 pounds + 0]/3 = 31 1/3 block pounds. The best solution is at Carrie's house.

Let's temporarily suppose that in preparation for their meeting, Marc will need to first visit a convenience store located 20 blocks west of his house (in a direction opposite of where his colleagues live) to purchase the liquid refreshment. He can take a bus to the store, but must walk (carrying the load) to the meeting place. With this information, does the location minimizing average weighted walking distance remain at Carrie's house, or is the best location now west of her house? The answer is that Carrie can still remain at home while Marc and Tej (poor Marc!) will need to walk to her house. In fact no matter how far west Marc lives relative to Carrie (2 blocks or 20 miles), the meeting point that minimizes the average effort remains at Carrie's home

So what is the insight into this problem? The answer lies in Fig. 6.2.

To make the situation a little more interesting, let us suppose that Carrie can leave some of the materials at home so the weight she must carry is now only 4 pounds. Figure 6.2 shows the home locations of Marc, Tej, and Carrie. In addition, the figure shows a cone pointed at each student's home location. The cones represent individual weighted distance functions. Thus, for example, considering the location at 301 Main street, Marc's weighted distance to that location is 2 blocks × 3 pounds = 6 block-pounds

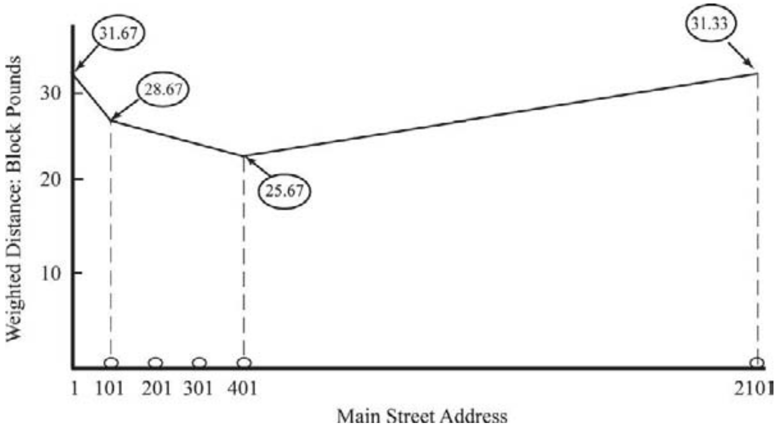


**Fig. 6.3** The average distance traveled for each meeting place

while Tej's is 1 block x 2 pounds = 2 block-pounds. (Even though she has reduced her load to 4 pounds, Carrie's weighted distance cone is too steep to show its value at 301 Main Street.) Since the overall objective of the location problem is to minimize average weighted distance, we need to create a figure that represents the sum (at any location) of the values obtained from each of the cones and subsequently divided by 3. Figure 6.3 illustrates this construction. The circled numbers on the figure are the average weighted distance values at 1, 101, 401, and 2101 Main Street. Note that with Carrie's reduced load to 4 pounds, the best location has shifted back to Tej's house. (Maybe Carrie should have insisted on needing all of the books!)

The principle insight is that the weighted average distance function is always bowl shaped (piecewise linear and convex) and the slope of this function only changes at the locations of travel origin (the student's homes). The optimal location is characterized by a point where the slope of this function changes from negative to positive. This point is at a *median point* and is where no more than 1/2 of the total weight lies strictly to the left (the right) of the point. Thus with the students carrying materials to the meeting location, and Carrie with 4 pounds, the total weight is 3 + 2 + 4 = 9 pounds. Note that 3 pounds of total weight lies strictly to the left of Tej's home so that 3 ≤ 9/2, while 4 pounds lies strictly to the right where 4 ≤ 9/2. When Carrie must carry 6 pounds, the total weight is 11 pounds and so 3 + 2 ≤ 11/2 pounds are strictly to the left of her home while there is no weight strictly to the right. Thus in this instance, Carrie's home would be the best location.

Another important insight in the solution to these location problems is that it is not dependent on the actual magnitude of the distances. Suppose we denote the location of Marc by 1, the location of Tej by 1 + $x$, and the location of Carrie by 1 + $x$ + $y$ where $x > 0$ and $y > 0$. The median property says that the optimal location does not change as the values of $x$ and $y$ change. The value of $x$ can be 3 blocks or 30 blocks and the value of y can be 17 blocks or 37 million blocks—the solution does not change. What does change is the total effort by Carrie, Tej, and Marc.

## Problem Formulation and Solution

To build structure for the above arguments, let Main street be represented by the real line with $a_1 = 0$ = Marc's house, $a_2 = 3$ (3 blocks from Marc) = Tej's house, and $a_3 = 20$ = Carrie's house. Letting $x$ represent the location for the meeting place, the objective is to find $x$ to minimize

$$[3 \mid x{-}a_1 \mid +2 \mid x{-}a_2 \mid +4 \mid x{-}a_3 \mid]/3 = [3 \mid x{-}0 \mid +2 \mid x{-}3 \mid +4 \mid x{-}20 \mid]/3,$$

where the coefficients are the weights in pounds the various students must carry. Note that $\mid x - a_j \mid$ measures the distance student j must walk.

If there are $n$ students located at $a_1, a_2, \ldots, a_n$ and the amount of weight that student $i$ has to carry is $w_i$ then the problem is to find $x$ such that it minimizes

$$w_1 \mid x - a_1 \mid + w_2 \mid x - a_2 \mid + \ldots. + w_n \mid x - a_n \mid \; = \sum_{i=1}^{n} w_i \mid x - a_i \mid.$$

The optimal $x$ satisfies two properties. First, it is at one of the existing locations $(a_1,\ldots,a_n)$ and second, it is the median point. Suppose $W$(i) denotes the sum of the weight at $i$ and all weights to the left of location $i$, and $W$ denotes the sum of all the weights. The solution is then: starting from point 1, find the first point $j$ such that $W(j) > = W/2$. In the meeting place problem note that $W = 9$, $W(1) = 3$, $W(2) = 3 + 2 = 5$ and $W(3) = 3 + 2 + 4 = 9$. $W(2)$ satisfies the median condition since $W(1) < 9/2$ and $W(2) > 9/2$.

## Tej Relocates

Suppose that instead of living at 401 Main Street, Tej lives 6 blocks north of that location. Given that Marc, Tej and Carrie continue to carry 3, 2, and 4 pounds of materials, where would be the best location for a meeting place? Would the best location be at Tej's new home? To answer this question, we first recognize that we are now faced with a location problem in two dimensions instead of one. Thus we need to decide on "walking paths" the students will take. Given that the problem is posed in an urban area, we will assume that the students will confine their travel to sidewalks, and so movement will be either in an east/west direction or in a north/south direction. In fact, if a student wishes to travel between two points with coordinates $(a_1, b_1)$ and $(a_2, b_2)$, the distance traveled will be $|a_1 - a_2| + |b_1 - b_2|$. Thus with Marc's home located at (0, 0), Tej's at (3, 6) and Carrie's at (20, 0), the problem is to find a point $(x^*, y^*)$ that minimizes

$$\{3[\mid x - 0 \mid + \mid y - 0\mid\} + 2\{\mid x - 3 \mid + \mid y - 6\mid] + 4[\mid x - 20 \mid + \mid y - 0\mid]\}/3.$$

Note that the problem separates into two distance problems: a problem in *x* and another in *y*. Thus to find ($x$*, $y$*) we simply find $x$* that minimizes

$$[3|x - 0| + 2|x - 3| + 4|x - 20|]/3,$$

and $y$* that minimizes

$$[3|y - 0| + 2|y - 6| + 4|y - 0|]/3.$$

But, each of these problems is basically a problem in one dimension and we know that a median point solves each such problem. Thus in the *x* dimension, a median occurs at $x$* = 3. In the *y* dimension, a median occurs at $y$* = 0. (Note that $y = 0$ is a median since a weight of $2 \le 9/2$ lies strictly above $y = 0$, while no weight lies strictly below $y = 0$.

With this information, we now know that the best meeting place will be at ($x$*, $y$*) = (3, 0), the location of Tej's previous home! Thus we have the interesting fact that even though the best *x* (best *y*) coordinate will be coincident with an *x* (*y*) coordinate of one of the demand points, the best location in two dimensions may be at a point that is not coincident with any demand point. We note that this fact is true in dimensions higher than two. Consider any dimension and arrange the weights on this axis as per their coordinates on this dimension. The optimum coordinate for this dimension would satisfy the median condition with respect to the weights thus arranged. In particular, given *n* points in *k*-dimensional space, where point i has coordinates $(a^i_1, a^i_2, \ldots, a^i_k)$ and weight $w_i$, $i = 1,.., n$, and where the objective is to find a point $X^* = (x_1^*, x_2^*, \ldots, x_k^*)$ that minimizes

$$\sum_{i=1}^{n} w_i \Big( \sum_{i=1}^{k} |a^i_j - x_j| \Big),$$

an optimizing point will have the property that the *j*th coordinate will be a median point with respect to the weighted *j*th coordinates of the demand points. That is, $X$* will have the property that

$$\sum_{i:a^i_j < x^*_j} w_i \le \left\{ \sum_{i=1}^{n} w_i \right\} / 2, \quad \text{and}$$

$$\sum_{i:a^i_j > x^*_j} w_i \le \left\{ \sum_{i=1}^{n} w_i \right\} / 2.$$

## What About Multiple Locations?

Let us return to the problem faced by the three students, but add a new aspect to it. Recall that Marc lives at (0, 0) and carries a weight of 3 pounds, while Tej and Carrie live at (3, 6) and (20, 0) carrying weights of 2 and 4 pounds respectively.

Carrie's materials include books as well as reports, and Tej will be bringing a computer that actually has a scanning feature. Carrie and Tej have decided that they should meet somewhere together first and scan 2 pounds of Carrie's materials (the reports), thus reducing Carrie's load since the scanned materials can then be discarded at this initial meeting site. Carrie and Tej can then proceed together to meet Marc, with each carrying 2 pounds. Thus the students must determine two meeting locations: $(x_1, y_1)$, where Carrie and Tej will meet, and $(x_2, y_2)$ where all three students will meet (with Carrie and Tej walking from $[x_1, y_1]$ to $[x_2, y_2]$). A formal statement of their problem is to find $(x_1{}^*, y_1{}^*)$ and $(x_2{}^*, y_2{}^*)$ that minimizes

$$\{4[|x_1 - 20| + |y_1 - 0|] + 2[|x_1 - 3| + |y_1 - 6|] + (2 + 2)[|x_1 - x_2| + |y_1 - y_2|] \\ + 3[|x_2 - 0| + |y_2 - 0|]\}/3.$$

Dividing by three reflects the fact that we are still concerned with average weighted travel distance for the students. The first and second terms are Carrie's and Tej's weighted travel distances to the location where scanning will occur. The third term is the weighted distance that Carrie and Tej must travel from the scanning location to the site where they will meet Marc. The fourth term is Marc's weighted travel distance.

Although the problem above cannot be solved in closed form, critical insights are still possible. Note that, for example, if the location of the three-student meeting site, $(x_2, y_2)$, is known, then the location of the scanning site $(x_1, y_1)$ can be found via the median conditions. That is, in the third term we can think of the point $(x_2, y_2)$ as a "demand point" with weight = 4. In this case the fourth term is a constant with no impact on the best location for the scanning site. Similarly, if $(x_1, y_1)$ is known, the best $(x_2, y_2)$ could be found via the median conditions.

These observations may suggest an iterative procedure to find the two locations, but that is not an efficient means for solving the problem. A much better alternative is to use linear programming by substituting linear expressions for the absolute value terms and using additional variables. Details on this conversion can be found in Francis et al. (1992).

**Building Intuition** The mean value is easily affected by very high or very low numbers in a set of numbers and is therefore not an appropriate measure of the central value of the set.

The median of a set of numbers (arranged in increasing order) is the middle value of the list and this is not affected by extreme values.

For certain location problems on a plane as well as on tree networks the median condition specifies the optimal location that minimizes the total weighted distance traveled to the new location.

A problem related to the above is popularly known as the *p*-median problem. For this problem, *p* servers (new facilities) must be located to provide service to a number of demand points. The objective is to locate the *p* servers to minimize the average weighted travel distance incurred (averaged over all of the demand points) when each travels to a closest new facility. Thus the key difference between the problem we study and the *p*-median problem is that in the latter problem the service "connections" are not known until the new facility locations have been determined. For one of the earliest heuristic methods for finding solutions to this difficult problem see Cooper (1964).

## Location on a Transport Network

Another example of a median location can be found in the problem of locating a single facility on a transport network where the objective is to find the location that minimizes the sum (or average of) weighted distances of several demand points to the facility. Roads, rail, and river networks are examples of such transport networks. Consider a special network, called a tree network, where there are no cycles in the network and there is a single path of travel between any two points on the network. A good example of a tree network is a river and its tributaries. Without loss of generality, we can suppose that each demand point is located at a node of the network. Thus if a demand point is not at a natural "branching point," we can declare the location to be a trivial node. In Fig. 6.4, the filled circles are the demand points and the hollow circles are the trivial nodes. Also, it is unnecessary that every branching point be the location of a demand point. We will call the demand at a node as the weight of the node. A useful result is that an optimal solution can be found by only considering nodes of the network. This result also holds for location problems on networks more general than tree networks (see Hakimi 1964)

The following mnemonic rhyme (Hua Lo-Keng et al. *1962*) provides a procedure to this problem. The second and the third verse also consider the case when

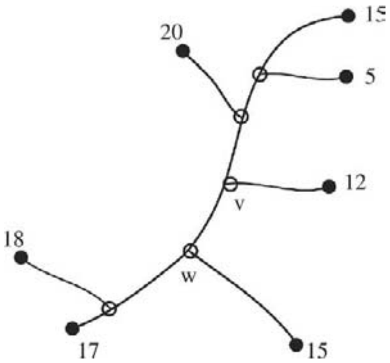


**Fig. 6.4** A tree network with demand nodes and trivial nodes

the network has loops. We, however, advise that the time required to find a solution using this method (when loops are present) will rise exponentially with the number of loops in the network.

*When the routes have no loops,*
*Take all the ends into consideration,*
*The smallest advances one station.*
*When the routes do have loops,*
*A branch is dropped from each one,*
*Until there are no loops,*
*Then calculations as before are done.*
*There are many ways of dropping branches,*
*The calculation for each must be assessed,*
*After figuring all, we then compare,*
*And break the loop in the case which is best.*

For the tree network, the method proceeds by collapsing the node with the smallest weight. The node with the smallest weight is removed and its weight is added to the node it was connected to. On a tree network, following this process will eventually lead to a node called the median node. The median node satisfies the property that no more than one-half of the total weight can be in any portion of the subtree connected to it by a link. In Fig. 6.4, the node with weight 5 will be the first node to be collapsed followed by the node with weight 12. The resulting tree after these two steps is shown in Fig. 6.5. Following the algorithm we will eventually get the arc (*v*, *w*) with weight 52 on node *v* and weight 50 on node *w*. In the last step, node *w* will be collapsed into node *v*, which is the optimal location.

More formally, let $a_i$, $i = 1,..,n$ be the locations of the *n* demand points, $w_i$ be the weight associated with demand point *i*, and *x* be the location of the facility. The objective is to find $x^*$ on the network that minimizes

$$\sum_{i=1}^{n} w_i d\left(a_i, x\right).$$

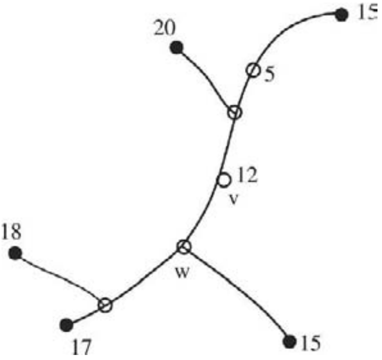


**Fig. 6.5** The tree network after collapsing two nodes

where d($a_i$, $x$) is the distance of travel on the network between demand point $i$ and $x$.

Then, for any node $v_j$ of the tree, the branches defined by $v_j$ are the resulting pieces, say $\{T^k(v_j), k = 1,\ldots\}$ of the original tree found by removing node $v_j$. Then letting $W^k(v_j)$ be the sum of the weights of all demand points in $T^k(v_j)$, and $W$ = the sum of all demand points in the tree, an optimal solution to the location problem will be a node $x^*$ where $W^k(x^*) \leq W/2$ for all branches $T^k(x^*)$. That is, no more than one half of the total weight can be in any branch defined by $x^*$. In Fig. 6.4, the total weight of all the nodes is 102. The three subtrees of the optimal location, node $v$, have weights of 50, 40, and 12, respectively.

The insight associated with the solution to this problem is similar to our first example. Note that if some node $v_j$ has a branch where $W^k(v_j) > W/2$, then movement into branch $T^k(v_j)$ from $v_j$ will decrease the objective function value. Once again the actual distances on the network do not affect the optimal location that minimizes the sum of the weighted distances. Thus even after stretching or shortening the links of the tree if the procedure is reapplied, the optimal location will remain unchanged.

## When Uncertainty is Involved

We now consider a problem that is more of a planning issue, with an additional wrinkle that events are no longer fully known. That is, suppose we want to plan for a future when we do not what will happen. This situation is unlike our meeting location problem where we knew exactly how many students are meeting, their current address and the load they will carry.

Avanti, the chef and owner of the Maharaja Restaurant is contemplating the quantity of chicken to order from her supplier for Sunday dinner, typically the busiest night at the restaurant. In the highly competitive restaurant business, controlling the cost of food is crucial. For the past 2 years, Avanti has been meticulously keeping data on her requirements for various ingredients that she uses in her restaurant. As a result of her hard work, she has the data but she is not quite sure how to use it to decide what she should order. For chicken, she knows that the demand has fluctuated between 50 lbs and 80 lbs. The table below shows the relative frequency of her need for various quantities of chicken for Sunday dinner. These frequencies (given in fractions) tell us how likely the demand for chicken will be on this coming Sunday. For example, the probability that the demand for chicken will be 50 pounds is 0.1 or 10%. The cost of fresh chicken from Avanti's supplier is $3 per pound. The restaurant is closed on Monday, so any chicken that is not used by closing time on Sunday is discarded. In case she finds the demand for chicken exceeds what she had ordered, she uses her stock of frozen gourmet chicken to make up the difference. Gourmet chicken is of very high quality and costs $6 per pound, twice the $3 per pound that she pays for the fresh chicken.

| Demand (lbs) | 50 | 55 | 60 | 65 | 70 | 75 | 80 |
|---|---|---|---|---|---|---|---|
| Probability, $P$ | 0.10 | 0.20 | 0.25 | 0.20 | 0.15 | 0.05 | 0.05 |

Note that if Avanti orders 60 lbs and demand is for 50 lbs, she will waste 10 lbs of chicken, and the cost of this mismatch is \$3*(60−50) = \$30. On the other hand, if the demand is 70 lbs then she has to use 10 lbs of frozen chicken, which would cost her an additional \$3*(70−60) = \$30. Thus the cost of ordering a pound more than the demand is the same as the cost of ordering a pound less than the demand. Since she does not know the exact demand for chicken on Sunday, the best Avanti can do is to minimize the cost she will incur in the long run. If she continues to order 60 lbs for every Sunday, the expected cost per week is $0.1*3*(60-50) + 0.2*3*(60-55) + 0.25*3*(60-60) + 0.2*3*(65-60) + 0.15*3*(70-60) + 0.05*3*(75-60) + 0.05*3*(80-60)$ or $\sum 3*P(D)*|D\text{-}60|$. Thus the correct order quantity is one that minimizes $3*\sum P(D)*|D-Q| = 3*\min \mathrm{E}(|D-\mathrm{Q}|)$. However, the value 3 has no effect on the optimal value of $Q$ since it is a constant in the minimizing process. Thus, we wish to find the value of $Q$ that minimizes $\mathrm{E}(|D\text{-}Q|)$. It is well-known that the minimizing value of $Q$ is the median of the distribution of $D$. In Avanti's case, the median is 60 pounds of chicken since $\mathrm{P}(D < 60) \le 0.5$, and $\mathrm{P}(D > 60) \le 0.5$.

As noted above, the solution to the problem is at the median of the distribution of demand as long as the unit cost of ordering more than the demand is the same as the unit cost of ordering less than the demand. Avanti's problem is actually an instance of the Newsvendor Problem which is discussed in Chap. 7. In Avanti's case, the cost of over- and under- ordering are the same, but often these two costs are not the same and so the median of the demand distribution may not necessarily be the optimal choice. However, a "median type" solution does solve the Newsvendor Problem as we discuss next.

## Medians and the Newsvendor Problem

As discussed in Chap. 7 of this volume, with $P(D)$ as the distribution on demand $D$, and with $c_u$ and $c_o$ as the unit costs of understocking and overstocking respectively, the Newsvendor Problem is to choose the stocking quantity $Q$ that minimizes the expected cost of understocking and overstocking. With $(x - y)^+$ defined to be $\max\{x - y, 0\}$ we can express this expected cost as $N(Q) = \mathrm{E}\{c_{\mathrm{u}}(D - Q)^+ + c_{\mathrm{o}}(Q - D)^+\}$. With $D$ as the random variable, we can use the fact that $\mathrm{E}\{(D - Q)^+\} = \mathrm{E}\{[|D - Q| + \mathrm{E}(D) - Q]/2\}$. Upon substitution we obtain $N(Q) = ((c_{\mathrm{u}} + c_{\mathrm{o}})/2)\ \mathrm{E}\{|D - Q|\} + ((c_{\mathrm{u}} - c_{\mathrm{o}})/2)\ (\mathrm{E}(D) - Q)$.

Suppose the distribution on $D$ is discrete, and let $P_i$ be the probability that demand is $D_i$, $i = 1,\ldots,n$. Each demand quantity could be thought of as a point on the real line. The first term contributes a weight of $(c_{\mathrm{u}} + c_{\mathrm{o}})/2)\ P_i$ for each demand point $i$. Suppose $c_u > c_o$, Then the last term is decreasing in $Q$ and to account for this we add an additional weight of $(c_{\mathrm{u}} - c_{\mathrm{o}})/2$ to the weight of demand point $n$. However, if $c_o > c_u$, then the last term in the objective function is increasing in $Q$ and we add the weight $(c_{\mathrm{o}} - c_{\mathrm{u}})/2$ to the weight of first demand point.

Thus, to solve the Newsvendor Problem as a median problem, reformulate the problem as follows:

1. Compute $v_i$= ($c_u$ + $c_o$)/2) $P_i$, $i = 1,\ldots,n$.
2. If ($c_u$ - $c_o$)/2) > 0, set $w_i = v_i$, $i = 1,\ldots,n$-1, and $w_n = v_n$ + ($c_u$ - $c_o$)/2).
3. If ($c_u$ − $c_o$)/2) < 0, set $w_1 = v_1$ − ($c_u$ − $c_o$)/2), and $w_i = v_i$, $i = 2,\ldots,n$.
4. Identify the index $i$* where $\sum_{i<i^*} w_i \le \frac{1}{2}\sum_{i=1}^{n} w_i, and \sum_{i>i^*} w_i \le \frac{1}{2}\sum_{i=1}^{n} w_i$ , and set $Q^* = D_i^*$.

Thus, the Newsvendor Problem can be solved as a median problem by appropriately defining the weights.

Avanti's problem is a very special case of the above since $c_u = c_o$ = $3, and so the adjustments in steps 2 and 3 above are nil. But suppose the cost of the chicken goes up to $4.25 from $3. The new purchase cost changes the overstocking and understocking costs to $c_o$ = $4.25 and $c_u$ = $1.75. Since ($c_u$ − $c_o$)/2) < 0 we set the weights following steps 1 and 3 which are given below. The median condition is now satisfied by the demand level of 55—the new order quantity. Note that the amount ordered has reduced from 60 to 55 as the cost of over ordering has gone up and the cost of under ordering has reduced.

| Demand (lbs) | 50 | 55 | 60 | 65 | 70 | 75 | 80 |
|---|---|---|---|---|---|---|---|
| Probability | 0.10 | 0.20 | 0.25 | 0.20 | 0.15 | 0.05 | 0.05 |
| w | 1.55 | 0.6 | 0.75 | 0.6 | 0.45 | 0.15 | 0.15 |
| Σw | 1.55 | 2.15 | 2.9 | 3.5 | 3.95 | 4.1 | 4.25 |

## Applications

Our applications thus far are on using the median condition to solve certain location and inventory problems. The two-dimensional case is also applicable for location of equipment within a warehouse or a production facility. Another application of the median condition is in a product design problem. Suppose a product can be represented by a one dimensional attribute and the customers are characterized by an ideal level for this attribute. Customer segment $i$ has a population of $w_i$ and desire an ideal level of $a_i$ for the attribute. The disutility of customer segment $i$ from a product with attribute level $b$ is $w_i$*|$a_i$ − b|. The product designer wants to specify the attribute level such that it minimizes the average weighted disutility over all the customers. This optimal level of attribute satisfies the weighted median condition and can be found easily.

The term median is often used (or should be used) while reporting summary statistics and we discuss this next.

In June 2005, there were seven houses sold in CherryVille with prices (in thousands), of $80, $85, $85, $90, $95, $100, and $110. Thus the average price of the houses sold in June is $92.14K. In this instance, the median price, defined as the middle value of the list arranged in an increasing order, is $90K. The mean and the median prices are quite similar and tell us the typical price of the houses sold. Now let us change this data a bit. Suppose the most expensive house sold was not priced at $110K but was the house that belonged to the basketball coach and was sold for $550K. Now the average price of the houses sold is $155K. Obviously this price is higher than all but the coach's house and does not adequately represent the typical price. The median price, however, is still $90K.

Note that the median is defined in exactly as the same way as the location that defines the median condition in the student meeting location problem with all the weights equal to one. The median does not change if one of the values becomes large or small as long as the changing price is not the median and it remains on the "same side" of the original median price. Similarly the optimal location of a point that minimizes the sum of distances walked by each student does not change if one of the students starts from a different location as long as the point from where the student walks does not switch the side on which it is relative to the original median. Returning to the house price example, as long as the fourth largest price remains at $90K, the actual values of the remaining six houses do not matter as long as three of them are less than $90K while three are more than $90K. The idea behind the median is that about half of the houses were sold at prices that were higher than the median and the other half were sold at prices that were lower than the median. Similar insights were derived for the location problem.

Our point is that the median value is often a more representative summary statistic, even though the average is frequently used. Data such as "average sale per customer" is prone to similar distortions due to outliers, although companies frequently use this summary statistic. In reporting income, house prices, grades, height of citizens, etc. the median is more appropriately used; it is less prone to be affected by "outliers" and is more often apt to convey the typical value of the quantity of interest.

## Historical Background

Evidently, the use of the weighted median as a location to minimize the sum of weighted distances to fixed points on the real line has its beginnings in the field of statistics. F. Y. Edgeworth (see Bowley 1928) used (and proved) the result in his analysis of error associated with the problem of using multiple observations of an unknown quantity to make statements about the quantity in question. He argues that the median of the observations is the "best mean" in comparison with other possible choices such as the arithmetic mean or the geometric mean. Dodd (1938) made use of the weighted median result in a study of the properties of the "$r^{th}$ *k*-tile" of a set of numbers, $1 \leq r \leq k - 1$. Ekblom (1973) also used the result in a study of nonlinear median estimators in the case of an even number of data points. Eells (1930), in an attempt to correct misstatements regarding properties of the "center of population" appearing in the special bulletin of the 14 (1920) Census, noted that proof of the median conditions appears in several early books on statistics, e.g., Kelley (1923). Snyder (1971) makes use of the median conditions (in one dimension) to show that finding the point *(x*, y*)* that minimizes total travel, via a rectangular network of roads, from points of "relative frequency of accidents" distributed via the joint density function *f(x, y)* can be found by finding the median in the x coordinate and the y coordinate using the marginal density *f(x)* and *f(y)*, respectively. Rosenhead (1973) and Mole (1973) provide comments on Snyder's analysis including mention of the early work in statistics. Bindschedler and Moore (1961), and Wesolowski and Love (1971) are early papers that study the median problem in two dimensions with applications in manufacturing facility layout. Francis et al. (1992) study the median conditions in a location setting and also provide a means of constructing contour lines—loci of points about the minimizing point that have a fixed objective function value.

## References

Bindschedler, A. E. and J. M. Moore (1961) "Optimal Location of New Machines in Existing Plant Layouts," *Journal of Industrial Engineering*, **12**, 41.

Bowley, A. L. (1928) *F. Y. Edgeworth's Contributions to Mathematical Statistics*, A. M. Kelley Publishers, Clifton, NJ.

Cooper, L. (1964) "Heuristic Methods for Location Allocation Problems," *Siam Review*, **6**, 37–53.

Dodd, E. L. (1938) "Definitions and Properties of the Median, Quartiles, and Other Positional Means," *The American Mathematical Monthly*, **45**, 302–306.

Ekblom, H. (1973) "A Note on Nonlinear Median Estimators," *Journal of the American Statistical Association*, **68**, 431–432.

Eells, W. C. (1930) "A Mistaken Conception of the Center of Population," *Journal of the American Statistical Association*, **25**, 33–40.

Hakimi, S. L. (1964) "Optimal Locations of Switching Centers and the Absolute Centers and Medians of a Graph," *Operations Research*, **12**, 450–459.

Hua Lo-Keng et al. (1962) "Application of Mathematical Methods to Wheat Harvesting," *Chinese Math*, **2**, 77–91.

Kelley, T. L. (1923) *Statistical Method*, Macmillan, New York.

Snyder, R. D. (1971) "A Note on the Location of Depots," *Management Science*, **18**, 97.

Rosenhead, J. (1973) "Some Comments on "A Note on the Location of Depots," *Management Science*, **19**, 831.

Mole, R. H. (1973) "Comments on "A Note on the Location of Depots," *Management Science*, **19**, 832.

Francis, R. L., L. F. McGinnis, and J. A. White (1992) *Facility Layout and Location: An Analytical Approach*, Prentice-Hall, Englewood Cliffs, NJ.

Wesolowski, G. O. and R. F. Love (1971) *Operations Research*, **19**, No. 1. 124–130.

# Chapter 7
# The Newsvendor Problem

**Evan L. Porteus**
**Stanford University**

*The newsvendor problem has numerous applications for decision making in manufacturing and service industries as well as decision making by individuals. It occurs whenever the amount needed of a given resource is random, a decision must be made regarding the amount of the resource to have available prior to finding out how much is needed, and the economic consequences of having "too much" and "too little" are known.*

## Introduction

Tyler has just been put in charge of deciding how many programs to order for this week's home football game between his team, the Rainbows, and the visiting team, the Warriors. Each program ordered costs \$1.25 and sells for \$5.00 each. He found out from Nar Lee, the experienced department administrator, that 8,500 programs were ordered for the last home game and only 7,756 were sold. He thinks to himself that his problem is solved. He'll simply order 7,756 programs. But he decides to discuss the issue with his friend, Shenn, who is a little bit nerdy, but tends to tease people she likes.

As Tyler is explaining his thinking, Shenn cuts in. "The last home game wasn't that big a deal. We were playing against the Titans and they had a pretty pathetic record coming into the game. Remember, it rained a bit before the game, too, which probably kept some people at home. This game's going to be a biggie, with strong implications for each teams' chances for a postseason bowl game. Don't you think it's much more likely that you could sell more programs for this game?" Tyler agreed but felt overwhelmed with having to figure out how many more. Shenn helped him understand the difference between the unit *demand* for programs and unit *sales*: The (unit) demand consists of the number of programs that would be sold if he had an unlimited number available, whereas (unit) sales are limited by the number of programs ordered. "Suppose you order 8,000 programs. If demand turns out to be for 9,000 programs, then you will sell all 8,000 that you ordered, but no more. However, if demand is for 7,000 programs, then sales will also equal 7,000 and you will have 1,000 unsold programs."

D. Chhajed and T.J. Lowe (eds.) *Building Intuition*: *Insights From Basic Operations Management Models and Principles*.
doi:10.1007/978-0-387-73699-0, © Springer Science + Business Media, LLC 2008

"I get it. But I don't want to have any leftovers. That will make me look bad. I still like that 7,756 number. I'm almost sure to sell them all. They sell for $5.00 and cost only $1.25, so I'd get the full $3.75 for each of them."

Shenn looked like she was settling in for a long discussion. "Have you thought about trying to look good, rather than avoiding looking bad? If you could increase the net return from selling programs, the proceeds would be mightily appreciated by the department. I think that what my econ prof called *opportunity cost* might be useful here. Suppose you decide to order 8,000 programs. It's certainly true that if you only sell 7,000, then you will have 1,000 unsold programs that cost $1.25 each, for a total of $1,250 and if you had ordered only 7,000 you would have saved that $1,250, without changing how many you sold. However, if demand is for 9,000 programs, then you will only sell 8,000, and if you had ordered 1,000 more, you would have sold them all, for a net profit of $3.75 each, representing a total gain of $3,750. The $3,750 opportunity cost of having 1,000 too few is much more than the $1,250 opportunity cost of having 1,000 too many. It may make sense to err on the side of ordering too much rather than too little. What do you think?" Tyler had to agree, but he could feel a headache coming on. "Let's go talk to Nar about how to think about the possible demand quantities for this game."

It did not take Nar long to power up his PC and print out a histogram of program sales for the past 3 years.

Tyler was the first to comment on the chart, "Well, I can see that we've never sold more than 9,500 programs, so I don't have to worry about ordering more than that!"

Shenn jumped in, "Not so fast, Bubba-head. You're forgetting the difference between demand and sales. When we sold 9,500 programs, that is all we ordered. We likely would have sold more if we had them. About all we can say is that the demand will be random. Let's figure out how many to order for an example. We can then go back and fine-tune our assumptions. Let's assume that there are only five possible outcomes, 7,000, 8,000, 9,000, 10,000, and 11,000, with probabilities of 0.1, 0.2, 0.4, 0.2, and 0.1, respectively."

Tyler exclaimed, "I like the idea of ordering 7,000 programs, then. We get the full $3.75 margin on each, for a total return of $26,250, without any unsold programs." He was also thinking to himself that he didn't even need to remember what probabilities meant to reach this conclusion.

"But you might be able to do better if you order more. The usual way to evaluate a decision like this is to determine the *expected value* of the resulting return. We get it by determining the net profit for each possible outcome, multiplying it by the probability of that outcome, and then adding up over all outcomes. If I remember what my stats prof said, this measure corresponds to the *average* net profit that would be obtained per football game if exactly the same problem were faced many times. Let's see what we would get if we ordered 8,000 programs." It didn't take long to create the worksheet shown in Exhibit 7.1.

Tyler felt his headache reappear. Shenn explained, "This says that ordering 8,000 yields an expected net return of $29,500, a full $3,250 more than the $26,250 from ordering 7,000. If demand is 8,000 or more, we sell all 8,000, for a net return of

**Exhibit 7.1.** Analysis of 8,000

| Demand outcome | Probability | ($) Net revenue | Product |
|---|---|---|---|
| 7,000 | 0.1 | 25,000 | 2,500 |
| 8,000 | 0.2 | 30,000 | 6,000 |
| 9,000 | 0.4 | 30,000 | 12,000 |
| 10,000 | 0.2 | 30,000 | 6,000 |
| 11,000 | 0.1 | 30,000 | 3,000 |
| Total | 1.0 | | 29,500 |

**Exhibit 7.2.** Analysis of 9,000

| Demand outcome | Probability | ($) Net revenue | Product |
|---|---|---|---|
| 7,000 | 0.1 | 23,750 | 2,375 |
| 8,000 | 0.2 | 28,750 | 5,750 |
| 9,000 | 0.4 | 33,750 | 13,500 |
| 10,000 | 0.2 | 33,750 | 6,750 |
| 11,000 | 0.1 | 33,750 | 3,375 |
| | | | 31,750 |

$30,000, which is $3,750 more than we get if we order only 7,000. And the probability of that happening is 0.9. The only time we get less is if demand is for only 7,000. We get $1,250 less in that case, but the probability of that happening is only 0.1. That is, we have a 90% chance of getting $3,750 more and a 10% chance of getting $1,250 less. The expected increase is 0.9(3,750) – 0.1(1,250) = 3,250. Looks to me like ordering 8,000 is quite a bit better than ordering 7,000. Let's see what we get if we order 9,000" (Exhibit 7.2).

Tyler's enthusiasm began to grow, "Let's see. If only 7,000 programs are demanded, then we net $3.75 on each of them, for $26,250 and lose $1.25 on each of the 2,000 we bought but didn't sell, for a net return of $ 26,250 – $ 2,500 = $23,750. We get $5,000 more if demand is 8,000 instead of 7,000, because we sell each of them for $5.00 and haven't bought any more. We get another $5,000 if demand increases by another 1000 to 9000. And, …, aha! If demand increases yet another 1000 to 10,000, we don't get any more, because we only have 9,000 to sell. This says our expected return is $31,750, which is $2,250 more than when we order 8,000. Can that be right?"

"Absolutely. If demand is 9,000 or more, we get $3,750 more than with ordering 8,000, and the probability of that is 0.7. If demand is 8,000 or less, we get $1,250 less, and the probability of that is 0.3. The expected gain is 0.7(3,750) – 0.3(1,250) = 2,250."

Tyler felt a bell ring in his head, "Wait, wait. Let me guess. If we order 10,000 instead of 9,000, then the expected gain will be 0.3(3,750) – 0.7(1,250) = 250, because the probability of getting the $3,750 more is the probability that demand will be 10,000 or more, which is 0.3. I can't believe that we should order that many!"

"Get used to it, baby. What about ordering 11,000?"

"Well, let's see. The expected gain will be 0.1(3,750) – 0.9(1,250) = – 750, so we would lose by making that change. We should stay at 10,000 programs. Yay!"

## *What is Tyler's Newsvendor Problem?*

Tyler's problem is an example of what is often called the newsvendor problem. Its name derives from the context of a newsvendor purchasing newspapers to sell before knowing how many will be demanded that day. In the simplest version of the newsvendor problem, such as Tyler's, the opportunity costs of an overage or underage are proportional to the size of that overage or underage, with unit costs of $c_O$ and $c_U$, respectively. For example, Tyler's unit overage cost $c_O$ is $1.25: It is the additional return that would be received if, in the event that too *many* programs were ordered, one *fewer* program was ordered. Tyler would have saved the $1.25 unit cost for each program that was not needed. Similarly, Tyler's unit underage cost $c_U$ is $3.75: It is the additional return that would be received if, in the event that too *few* programs were ordered, one *more* program was ordered. Tyler would have sold each program that was demanded but unavailable for $5.00. He would have had to pay the $1.25 unit cost for each of them, netting a margin of $3.75 per program. We will discuss more general versions of the newsvendor problem at the end of this chapter.

As far as Tyler is concerned, he is happy with the solution to his problem. However, we want to build a model that represents a general form of his problem and see if we can find a systematic way to solve it. We also would like to see if there are any insights we can glean from the analysis that can aid our intuition.

**Building Intuition** Tyler's decision of how many programs to order requires evaluating the *tradeoff* between having too many and too few. There is something good and bad for each outcome. If he has too many, he has an *overage*. The bad part is that he has leftover unsold programs, which he bought but are now worthless. The good part is that every customer is satisfied (able to buy a program when desired) and, hence, there are no shortages. If he has too few, he has an *underage*. The bad part is that some customers were unsatisfied (wanted to buy a program but were unable to because he ran out). He could have done better by buying more because he would have been able to sell them at a substantial margin. The good part is that he has no leftover, unsold programs.

A *newsvendor problem* has three defining characteristics: (1) there is a random amount needed of some resource, (2) a single quantity of that resource must be selected prior to observing how much is needed, and (3) all relevant economic consequences can be represented by known (opportunity) costs in terms of either the amount of overage or the underage. The optimal quantity of the resource to make available is easy to identify when the overage and underage costs are proportional to their size.

**Exhibit 7.3.** Probability distribution of demand

| Number in list, $i$ | Demand outcome $d_i$ | Probability<br>$p_i = P(D = d_i)$ | Cumulative probability<br>$P(D \le d_i)$ |
|---|---|---|---|
| 1 | 7,000 | 0.1 | 0.1 |
| 2 | 8,000 | 0.2 | 0.3 |
| 3 | 9,000 | 0.4 | 0.7 |
| 4 | 10,000 | 0.2 | 0.9 |
| 5 | 11,000 | 0.1 | 1.0 |

In the example that we have solved, the demand is *discrete*, which means that the demand outcomes can be put into a list, as in Exhibit 7.3. Each entry in the list is numbered, and we can therefore talk about *outcome* $i$, for $i = 1,2, ...,5$. It is useful to have symbols for the pertinent quantities in the model: Let $d_i$ denote the demand (quantity) for outcome $i$ and $p_i$ the probability of that outcome (for each $i$). Let $D$ denote the random demand that occurs. For example, the probability that demand equals 8,000 is 0.2. In symbols, $P(D = d_i) = p_i$ for each i = 1,2, ... ,5.

Let us assume henceforth that it is possible to receive a positive expected return. (We have already seen that this is true for Tyler's problem.)

When demand is discrete, then we only need to consider ordering quantities that can be demanded: It can be shown that if it is better for Tyler to order 7,001 programs than 7,000, then it is even better to order the next larger demand quantity, namely 8,000. (The expected gain by going from 7,000 to 7,001 will also be the expected gain by going from 7,001 to 7,002, and so on.)

The *optimal* amount to order is one that yields the highest expected return. The optimal amount can be determined by *marginal analysis*: We identify the expected gain of going to the next larger demand quantity from the current amount. If it is positive, we make the move, and repeat the process. In Tyler's example, if we go from 8,000 to 9,000, then the expected gain is

$$0.7(3{,}750) - 0.3(1{,}250) = 2{,}250.$$

In general, if we go from $d_{\mathrm{i}}$ to $d_{i+1}$, then the expected gain is

$$P(D \ge d_{i+1})(d_{i+1} - d_i)(c_U) - P(D \le d_i)(d_{i+1} - d_i)(c_o).$$

For example, if $i = 2$, then the expected gain is

$$\begin{aligned} &P(D \ge d_{i+1})(d_{i+1} - d_i)(c_U) - P(D \le d_i)(d_{i+1} - d_i)(c_o) \\ &= P(D \ge d_3)(d_3 - d_2)(c_U) - P(D \le d_2)(d_3 - d_2)(c_o) \\ &= 0.7(1{,}000)(3.75) - 0.3(1{,}000)(1.25) = 2{,}250, \end{aligned}$$

confirming what we already knew.

### Finding the Optimal Solution

It can be shown that we should go from $d_i$ to $d_{i+1}$ if

$$P(D \le d_i) \le \frac{c_U}{c_O + c_U}.$$

The ratio $\gamma = c_U/(c_O + c_U)$ is called the *critical fractile* which, in Tyler's case, equals $c_U/(c_O + c_U) = 3.75/(1.25 + 3.75) = 3.75/5 = 0.75$. That is, if the inequality above holds for a particular possible demand quantity, then it is better to stock more than that quantity. For example, the probability that demand is no more than 7,000 is 0.1, which is less than 0.75, so it's better to stock more than 7,000, namely at least 8,000. Similarly, it's better to stock more than 8,000 and also 9,000. However, the probability that demand is no more than 10,000 is 0.9, which is *more* than 0.75, so it is not optimal to go to 11,000, which we already knew.

It is nice that finding the optimal amount is simple: we just go down the list of cumulative probabilities, find the last one that is less than the critical fractile, and the optimal level (to stock) is the next larger demand outcome in the list. For example, going down the list of cumulative probabilities in the rightmost column of Exhibit 7.3, the last one that is less than 0.75 is 0.7, and the optimal level is the next larger demand outcome, namely, $d_4 = 10{,}000$.

There are other ways to find the optimal amount. For example, we can start at the bottom of the list of cumulative probabilities and find the smallest one in that list that is more than the critical fractile. That demand quantity is optimal.

If one of the cumulative probabilities equals the critical fractile, then there are *alternative optimal solutions*: The demand quantity whose cumulative equals the critical fractile is optimal and so is the next larger demand quantity: Each quantity yields the same expected return. (The marginal return from going to the next larger demand quantity is zero, so there is a tie for what is best.)

Now that we know how to find the optimal solution, we do not need to restrict our consideration to simple problems with only a few possible different demand quantities. Instead of only five different quantities, we could have hundreds or thousands, and it would still be easy to find the optimal solution.

### Continuous Demand Distributions

In practice, it is often convenient to use *continuous* (cumulative) probability distributions, such as the *normal distribution*, as representations for the probability distributions of demand. Technically, this requires thinking that it is possible for a fractional number of programs, such as 7,337.473, to be demanded. However, from a practical perspective, we are only *approximating* the demand distribution. We are building a model of a practical problem and virtually any model of a managerial situation makes simplifying assumptions to aid the analysis. The art of building and analyzing models in this regard is to develop an instinct for which simplifying assumptions are good ones, in the sense that they facilitate the analysis and do not mislead the results in any important way. Use of continuous distributions to approximate random quantities that are discrete is widely accepted as being a potentially good simplifying assumption.

When the demand distribution is continuous, it is also easy to determine the optimal quantity: We select the quantity $Q$ such that

$$P(D \leq Q) = \frac{c_U}{c_O + c_U} \cdot$$

## Back to Tyler's Problem

Tyler and Shenn returned their attention to Nar and, after lengthy discussion and some statistical analysis, decided that a better representation of the random demand for programs (for the coming game) is a normal distribution with mean of $\mu = 9{,}000$ and a standard deviation of $\sigma = 2{,}000$. Exhibit 7.4 illustrates the normal distribution and the critical fractile solution.

## Normal Distribution

When the demand distribution is a normal distribution, then we can find the optimal order quantity either by using tables, found in statistics books, or by using built-in functions of software packages, such as Excel spreadsheets. Exhibit 7.5 makes it easy not only to find the optimal order quantity but the resulting expected overage and underage costs. The first column gives the critical fractile, $\gamma$, and the second

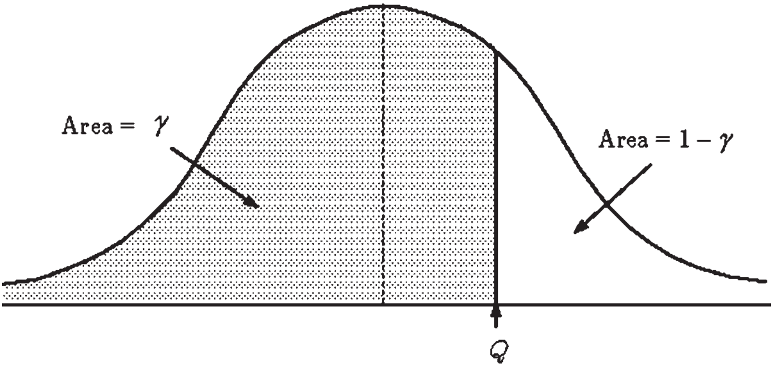
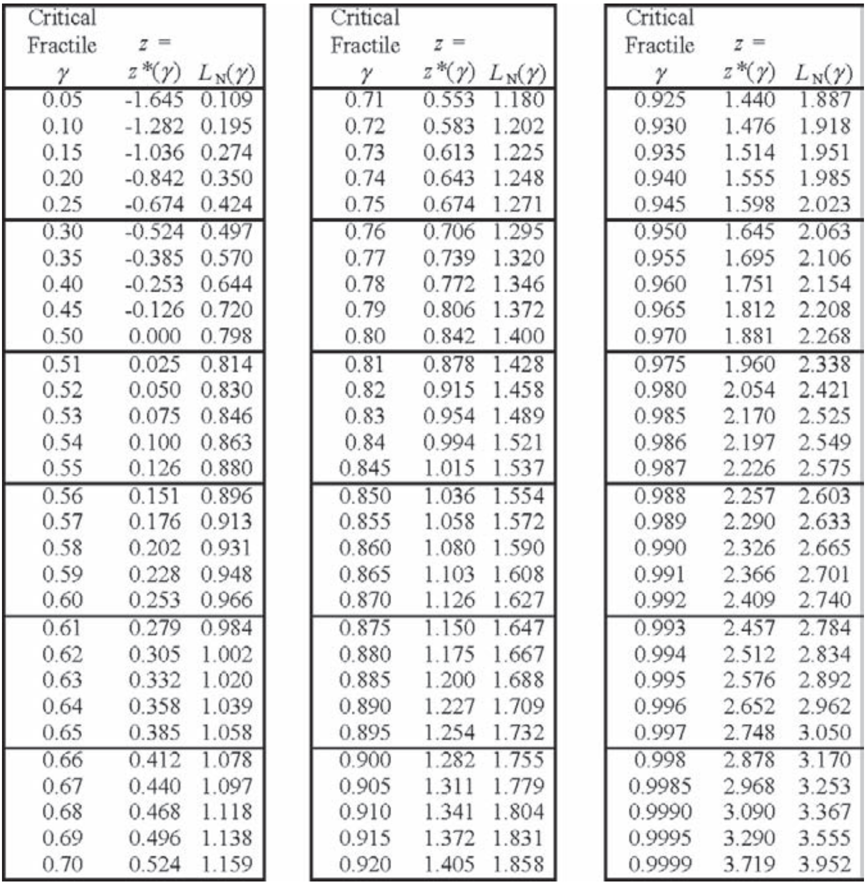


**Exhibit 7.4** Critical fractile solution



| Critical Fractile $\gamma$ | $z = z^*(\gamma)$ | $L_N(\gamma)$ |
|---|---|---|
| 0.05 | -1.645 | 0.109 |
| 0.10 | -1.282 | 0.195 |
| 0.15 | -1.036 | 0.274 |
| 0.20 | -0.842 | 0.350 |
| 0.25 | -0.674 | 0.424 |
| 0.30 | -0.524 | 0.497 |
| 0.35 | -0.385 | 0.570 |
| 0.40 | -0.253 | 0.644 |
| 0.45 | -0.126 | 0.720 |
| 0.50 | 0.000 | 0.798 |
| 0.51 | 0.025 | 0.814 |
| 0.52 | 0.050 | 0.830 |
| 0.53 | 0.075 | 0.846 |
| 0.54 | 0.100 | 0.863 |
| 0.55 | 0.126 | 0.880 |
| 0.56 | 0.151 | 0.896 |
| 0.57 | 0.176 | 0.913 |
| 0.58 | 0.202 | 0.931 |
| 0.59 | 0.228 | 0.948 |
| 0.60 | 0.253 | 0.966 |
| 0.61 | 0.279 | 0.984 |
| 0.62 | 0.305 | 1.002 |
| 0.63 | 0.332 | 1.020 |
| 0.64 | 0.358 | 1.039 |
| 0.65 | 0.385 | 1.058 |
| 0.66 | 0.412 | 1.078 |
| 0.67 | 0.440 | 1.097 |
| 0.68 | 0.468 | 1.118 |
| 0.69 | 0.496 | 1.138 |
| 0.70 | 0.524 | 1.159 |
| 0.71 | 0.553 | 1.180 |
| 0.72 | 0.583 | 1.202 |
| 0.73 | 0.613 | 1.225 |
| 0.74 | 0.643 | 1.248 |
| 0.75 | 0.674 | 1.271 |
| 0.76 | 0.706 | 1.295 |
| 0.77 | 0.739 | 1.320 |
| 0.78 | 0.772 | 1.346 |
| 0.79 | 0.806 | 1.372 |
| 0.80 | 0.842 | 1.400 |
| 0.81 | 0.878 | 1.428 |
| 0.82 | 0.915 | 1.458 |
| 0.83 | 0.954 | 1.489 |
| 0.84 | 0.994 | 1.521 |
| 0.845 | 1.015 | 1.537 |
| 0.850 | 1.036 | 1.554 |
| 0.855 | 1.058 | 1.572 |
| 0.860 | 1.080 | 1.590 |
| 0.865 | 1.103 | 1.608 |
| 0.870 | 1.126 | 1.627 |
| 0.875 | 1.150 | 1.647 |
| 0.880 | 1.175 | 1.667 |
| 0.885 | 1.200 | 1.688 |
| 0.890 | 1.227 | 1.709 |
| 0.895 | 1.254 | 1.732 |
| 0.900 | 1.282 | 1.755 |
| 0.905 | 1.311 | 1.779 |
| 0.910 | 1.341 | 1.804 |
| 0.915 | 1.372 | 1.831 |
| 0.920 | 1.405 | 1.858 |
| 0.925 | 1.440 | 1.887 |
| 0.930 | 1.476 | 1.918 |
| 0.935 | 1.514 | 1.951 |
| 0.940 | 1.555 | 1.985 |
| 0.945 | 1.598 | 2.023 |
| 0.950 | 1.645 | 2.063 |
| 0.955 | 1.695 | 2.106 |
| 0.960 | 1.751 | 2.154 |
| 0.965 | 1.812 | 2.208 |
| 0.970 | 1.881 | 2.268 |
| 0.975 | 1.960 | 2.338 |
| 0.980 | 2.054 | 2.421 |
| 0.985 | 2.170 | 2.525 |
| 0.986 | 2.197 | 2.549 |
| 0.987 | 2.226 | 2.575 |
| 0.988 | 2.257 | 2.603 |
| 0.989 | 2.290 | 2.633 |
| 0.990 | 2.326 | 2.665 |
| 0.991 | 2.366 | 2.701 |
| 0.992 | 2.409 | 2.740 |
| 0.993 | 2.457 | 2.784 |
| 0.994 | 2.512 | 2.834 |
| 0.995 | 2.576 | 2.892 |
| 0.996 | 2.652 | 2.962 |
| 0.997 | 2.748 | 3.050 |
| 0.998 | 2.878 | 3.170 |
| 0.9985 | 2.968 | 3.253 |
| 0.9990 | 3.090 | 3.367 |
| 0.9995 | 3.290 | 3.555 |
| 0.9999 | 3.719 | 3.952 |

**Exhibit 7.5** Newsvendor solution for the normal distribution

gives the stock level, expressed as the number $z^*(\gamma)$ of standard deviations above the mean. In Tyler's problem, $\gamma = 0.75$, so we see that $z^*(0.75) = 0.674$. Thus, the optimal stock level is

$$Q^* = \mu + z^*(\gamma)\,\sigma = 9{,}000 + 0.674(2{,}000) = 10{,}348.$$

The third column in Exhibit 7.5 shows the cost factor $L_N(\gamma)$, which gives the resulting expected overage and underage costs as follows:

$$c_O\,\sigma\,L_N(\gamma).$$

In Tyler's problem, we have

$$c_O\,\sigma\,L_N(\gamma) = 1.25(2{,}000)(1.271) = 3178,$$

which is the optimal expected opportunity cost. But Tyler is not satisfied with this number, because he can not relate it to the expected return, which is what counts. The formula for the expected return in Tyler's problem is

$$3.75\mu - c_O \sigma L_N(\gamma) = 3.75(9{,}000) - 1.25(2{,}000)(1.271) = 33{,}750 - 3{,}178 = 30{,}572,$$

which we will now explain.

The $3.75\mu$ is the expected return Tyler would get if he could defer deciding how many programs to order until after seeing how many programs would be demanded. In this case, for each quantity demanded, exactly that number would be ordered, and the full margin of $3.75 would be received on each of them. The resulting expected return would be 3.75 times the expected demand. This is sometimes called the *expected return under perfect information* because it corresponds to having a magic elf that has perfect foresight in that he can perfectly see ahead to what the random demand will be and whisper the amount to Tyler before he places his order. The expected overage and underage costs is also called the *expected value of perfect information (EVPI)* because it is how much better Tyler would do if he had the services of that magic elf: Because he must decide on how many programs to order before knowing what the random demand will be, the overage and underage costs would be eliminated if he had perfect information.

## *Insights*

There are useful insights that can aid your intuition when facing similar problems. They are easiest to understand when the probability distribution of demand is continuous, because in that case there is no chance that demand will exactly equal the amount ordered. Either $Q > D$ which means an overage occurs or $Q < D$, which means an underage occurs.

Both the unit overage and underage costs, $c_O$ and $c_U$, are strictly positive. (We will clarify below that the problem is improperly formulated otherwise.) The critical fractile $\gamma = c_U/(c_O + c_U)$, which is between zero and one, gives the *optimal overage probability*, $P(D < Q) = P(D \le Q)$, namely the optimal probability that you will have too much. Similarly, 1 - $\gamma$ is the optimal probability that you will have too little. The solution can be interpreted in gambling parlance: It is optimal to run odds of $c_U$ to $c_O$ of having too much, which means that the probability of having too much is $c_U/(c_U + c_O)$. Odds do not depend on their units, so we can also say it is optimal to run odds of $c_U/c_O$ to 1 of having too much. For example, if the (unit) underage cost is 3, 4, or 9 times the overage cost, then it is optimal to run a 75, 80, or 90% chance, respectively, of having too much. It is intuitive that the higher the relative cost of having too little is, the higher the optimal probability of having too much should be. The newsvendor model gives us more than that, namely, the exact way in which the probability should change.

One insight is that *never having leftover programs is bad, not good.* Tyler thought he would look good by not having any leftover, unsold programs, but our analysis shows that he should run a 75% chance (3–1 odds) of having leftovers. Tyler's boss needs to understand this, too, so he or she doesn't implicitly put pressure on Tyler to order too few and thereby receive lower expected returns.

A related insight is that *never having unsatisfied customers* (who tried to buy a program but were unable to, because the programs were sold out) *is also bad*. Tyler should run a 25% chance of having unsatisfied customers.

## Appeasing Unsatisfied Customers

Tyler took a marketing class once and he remembered that the mantra in that class was never to have unsatisfied customers. He could feel the headache returning. Unsatisfied customers might decide that they would stop attending games in the future, or would not bother trying to buy a program at future games. They might have such a bad taste in their mouths that they decide not to donate any money to the athletic department. These all have negative financial consequences that were not considered in our previous analysis. If Tyler can assess the expected financial loss from each such unsatisfied customer, then he can redo the analysis with the new considerations. For example, suppose he determines that each unsatisfied customer will yield an expected loss, of the types discussed, of $7.50. Then the underage cost changes to become $c_U = 3.75 + 7.50 = 11.25$. In the event of an underage (programs selling out), then, for each additional program, not only would it get sold, yielding a margin of $3.75, but the expected shortage cost of $7.50 would be eliminated. The new critical fractile is $c_U/(c_O + c_U) = 11.25/12.50 = 0.90$, which means that more programs should be ordered, the optimal probability of having no unsatisfied customers is 90%, and the optimal probability of having unsatisfied customers has been cut down to 10% from 25%. The optimal opportunity costs increase and the optimal expected return decreases.

Shenn, who took the same marketing class with Tyler but paid more attention to it, suggests that the analysis should not be left there. If there is some way to appease the unsatisfied customers that costs less than $7.50, then the athletic department would be better off doing that. For example, suppose all unsatisfied customers would be so happy to receive a $2 voucher to be redeemed at any food and drink concession at the game that they would carry no lingering resentment about not being able to buy a program. Then the underage cost could be reduced to $c_U = 3.75 + 2 = 5.75$, Tyler would order fewer programs than under the $7.50 shortage cost, and the expected return would increase. There are a couple of practical problems that must be solved in cases like this: The first is determining exactly how much is needed to appease an unsatisfied customer. (The fact that this amount might be random (differ across different potential unsatisfied customers) is not a problem, as long as we can estimate the *expected* amount.) The second is to assure that only truly unsatisfied customers are given the appeasement. For example, Tyler does not want any unsatisfied customers telling their friends and family at the game that all they need to do is pretend they want to buy a program and they will receive a $2 voucher. (The expected shortage cost could be much more than $2 for each (truly) unsatisfied customer in that case.)

## *Comparative Statics (Sensitivity Analysis)*

In the previous discussion, we saw that as the underage cost increases, the critical fractile, and hence the optimal probability of an overage and the optimal order quantity, increases. As the overage cost increases, the opposite happens.

If either the overage or underage cost increases, then the optimal expected costs increase, and for situations such as Tyler's, the optimal expected return decreases. Of course, if either the underage or overage cost decreases, the opposite happens.

Suppose for the remainder of this section that the demand distribution is normal. If the critical fractile equals 0.5, then it is optimal to stock the mean demand (zero standard deviations above the mean). If the critical fractile is above 0.5, then it is optimal to stock more than the mean and if the standard deviation increases, then the optimal order quantity increases. However, if the critical fractile is below 0.5, then it is optimal to stock *less* than the mean and if the standard deviation increases, then the optimal order quantity *decreases*. Regardless of the value of the critical fractile, the optimal expected opportunity costs increase as the standard deviation increases. The intuition here is that the effect of more uncertainty (increasing the standard deviation) on the optimal order quantity depends on which is higher, the overage or underage cost, but the effect on the costs is always bad.

If the mean demand increases by one (and no other changes are made), then the optimal order quantity also increases by one, but the optimal expected opportunity costs remain the same. The expected return increases. Tyler has the incentive to increase the mean (demand) and reduce the standard deviation.

## *One Stocking Decision for a Perishable Product*

The newsvendor model applies to problems of determining a single order quantity when unsold units of the product ordered will be obsolete and cannot be offered for sale again. Tyler's problem fits this description because the programs ordered for one game cannot be legitimately offered as the program for any subsequent game. Newspapers and magazines are similar.

The newsvendor model can be useful for determining stock levels of perishable produce at a grocery store: For example, if a grocery store gets new lettuce every day and throws out the old, unsold lettuce (or gives it to a food bank or otherwise diverts the leftover unsold lettuce), then the grocery store faces a newsvendor problem each day for lettuce. If you are a customer of such a grocery store and that store almost never runs out of lettuce, then either the margin on lettuce is extremely high, or the grocer estimates the financial cost of a customer finding no lettuce to be extremely high, or both. Incidentally, it is heard on the street that grocery stores get a significant amount of their profits from their produce sections. The newsvendor model does not capture all of the dynamics of selling produce at a grocery store. For example, if the grocer sells out of one type of lettuce, a customer may buy a different type of lettuce, as a substitute. To the extent that this kind of substitution takes place, the effective cost of a shortage is less and, thus, the optimal stock level can be lower. (If there is a type of lettuce that is the heavy favorite to be selected as a substitute for other types when they are out of stock, the way vanilla ice cream is the favorite substitute for other flavors, then the optimal stock level for that type of lettuce may be higher than if the effect of substitution were not considered.)

### *One Stocking Decision for a Nonperishable Product*

The newsvendor model can be useful for determining how much to order of a product that requires a long lead time to produce and will be sold only over a relatively short season. In this case, there is still only one stocking decision. Much fashion-oriented clothing falls into this category: It takes many months to procure materials, manufacture the clothing, and transport it to the location/country where it will be sold. Unsold merchandise at the end of the season is generally sold to another organization at a deep discount. In contrast to restocking lettuce, which is done daily, this decision would be made once for the selling season (such as the Christmas season). The overage cost would be the unit materials, assembly, and transportation costs less end of season scrap value.

### *An Improper Formulation*

Something is wrong with the formulation of the above problem if the overage cost is negative: This would happen if the end of season scrap value is assessed as more than the sum of the unit materials, assembly, and transportation costs. In this case, it would be optimal to buy an infinite amount, and each of the (infinite) number of unsold goods would net more scrap value than what they cost, yielding an infinite total return. In actuality, it is likely that if there are only a few unsold units, then they can be scrapped at a reasonably high value, perhaps more than what they cost. As the number of unsold units increases, the average scrap value decreases. A model of this situation is discussed further at the end of this chapter.

### *Stocking Decisions Under Buy-Backs*

Although the newsvendor model is nominally designed for determining the stock level of a product that becomes obsolete by the end of the selling season, it can also be useful for determining the stock levels of a product that still retains a great deal of value at the end of the selling season. In particular, here we assume that the supplier offers *buy-backs*, which means that the supplier will buy back any unsold units. For example, Tyler may also be responsible for deciding how many cases of soda to order for the drink stands to sell during the game, and unsold cases can be returned to the supplier for a substantial credit. Suppose that each case of soda costs \$4 from the supplier including delivery and that every unsold case of soda can be returned for a net of \$3 per case (\$3.50 less \$0.50 for labor and transportation getting the soda back to the supplier). If each case sells for \$24, then the unit underage cost would be \$20. (In case of an underage, Tyler could have bought one more case for \$4, sold it for \$24, and netted \$20.) The unit overage cost would be \$1. (If there is an overage, Tyler could have bought one fewer case for \$4, which would therefore not be available to be returned for \$3, for a net gain of \$1.) The critical fractile is 20/(20 + 1) = 0.953, so Tyler should run a 95.3% chance of having unsold soda and a 4.7% chance of running out.

Shenn again offers an opportunity for improvement by suggesting that Tyler not return unsold cases of soda until the end of the football season, because, as long as the unsold cases can be stored safely (without deterioration) between games for less than \$1 each, then Tyler should do that. For example, if a case of soda can be put into storage and put back into position for sale at the next game for \$0.60, then that approach would save \$0.40 per case. The unit overage cost becomes \$0.60. (If there is an overage, he could have bought one fewer case for \$4. That benefit would be cancelled out because it would therefore not be available to be sold at the next game, and, thus, a new case would have to be purchased then for \$4 delivered. The benefit would come from not having to spend the \$0.60 to store it between games.) The critical fractile increases and the expected costs decrease. Shenn warns him that this logic is valid only when he is assured that all unsold cases would require replacing at the next game if they were not available. For example, if Tyler has 1000 cases of unsold soda at the end of one big game and the next game is a small one where only 800 cases would be purchased if there were no leftover soda, then 200 of the unsold cases did not require replacing at the next game and a different unit overage cost applies to them. The overage costs would not be proportional to the size of the overage. This case of nonlinear overage costs is discussed at the end of this chapter.

## Improper Formulations and Their Resolution

An interesting situation arises if there are no handling, transportation, or other transaction costs and cases of soda can be returned for their purchase price. An initial analysis would yield an overage cost of zero. (In the case of an overage, Tyler could have bought one fewer case for \$4, which would therefore not be available to be returned for \$4, for a net gain of \$0.) The critical fractile would be $c_U / c_U = 1$, which means that Tyler should run absolutely no risk of having a shortage. If demand is really normally distributed, then there is no finite number $Q$ such that $P(D \leq Q) = 1$, so Tyler would have to stock an infinite amount. This is obviously not a practical solution. Something is wrong with the formulation. One resolution is to find that there must be some cost that was left out of the analysis. For instance, Tyler would not have enough room to store all the cases of soda in the world at the stadium, so space would have to be rented, which would cost something and create a nonzero overage cost, and, therefore, a finite optimal stock level. There are many other possible resolutions. For example, the supplier would realize that it would be a bad idea to absorb the handling and transportation costs on tons of soda that almost surely would be returned. The supplier would be inclined to require Tyler to share in some of the cost, and even seek at least a small gain for all soda that is returned.

Another resolution of the improper formulation is to incorporate the opportunity cost of capital. For example, for some small expensive items, such as jewelry, the handling and transportation costs are very small compared to the cost of the item. The overage cost should also include any holding cost that is incurred for unsold items, which in this case, includes the interest costs of the money needed to purchase the items. Even if the unsold items can be sold back at the same price they were bought at, having those items for a period of time requires the commitment of capital, which could have been used for other productive purposes during that period.

## *Further Insights on Buy-Backs*

Tyler figures that the more the supplier pays for buy-backs, the better off he will be. That is true if nothing else changes. However, in situations such as campus bookstores, the supplier can take nearly all the profits of the business. Professors decide what textbook to require for their courses, only one publisher prints each book, and there is a published retail price, so the bookstore has no ability to change the textbook, source it from a different supplier, or charge a higher price. The publishers then charge the bookstore a price that is nearly as high as the retail price. They also offer a buy-back rate that is nearly as high as its wholesale price. This combination means that the bookstore will make a little bit of money selling the books that are demanded, and not run any risk of having expensive, unsold textbooks after the demand has been realized. The publishers make virtually all the money. Without the buy-back system, the bookstore would have a high overage cost, leading to a low critical fractile, and a high probability that some students would not be able to buy the book, an untenable campus outcome. The bookstore would be pressured to stock more books than is optimal, and doing so might lead to losing money on a regular basis. The bookstore might even have to be closed. All these negative outcomes would likely lead to the revelation that the publishers are taking almost all of the profits. Pressure might be created on the publishers to reduce their wholesale prices. By using buy-backs, they keep their situation off the radar screen.

Publishers often go one step farther, by requiring the bookstores to send only the covers of the unsold copies of the books they buy back. This practice suggests that the shipping and handling costs incurred in sending the book back to the publisher are more than the marginal cost of printing it in the first place. (The publisher finds it cheaper to print more copies in the first place rather than have the complete books sent back for redirection to another bookstore later.) The publishers insist on getting the covers sent back, because they do not want to offer the temptation to the bookstores of requesting reimbursement (buy-backs) for books that they actually sold. The publishers often even include a message to consumers inside the cover that the book is "illegal" if it does not include its cover. (The publishers don't want the bookstores to send the covers back, get reimbursed for them, and then sell what's left of the books to consumers.)

## *Multiple Replenishment Problem for a Nonperishable Product*

A useful way of understanding whether the newsvendor model is useful for determining replenishment stock levels of a nonperishable product over a sequence of selling periods is whether we can safely assume that each period's problem starts with zero initial inventory. For instance, if unsold units available at the beginning of the next period can be returned for the same price as new ones can be purchased, then we can conceptually think of returning any unsold units and then buying them back (at the same price) as needed to get to the optimal stock level for the next period.

Similar logic is valid even if unsold units cannot be returned. Suppose that the problem is *stationary*, which means that each period is alike in the sense of the demand and the costs and revenues, and has an *infinite horizon*, which means that there are an infinite number of periods. Then the optimal stock level will be the same in each period. Thus, any unsold units at the end of a period can be used to reduce the number that are purchased at the beginning of the next period. Inventory management in this case is called *one for one* replenishment: Each unit that is sold in a period is replenished at the beginning of the next period. The overage cost is the holding cost, which includes both *financial holding costs*, which include interest costs, and *physical holding costs*, which include handling, storage, temperature control, insurance, and other costs incurred for each unit stocked for the period. The newsvendor model gives the optimal stock level. (See Chap. 9 in this volume for a discussion and use of the same idea.)

## *Capacity Management*

The newsvendor model can be useful in determining decisions other than stock levels. For example, Tyler's uncle, Rusty, works for an automobile manufacturer that must decide on how much *capacity* to have to manufacture each of its models for the approaching model year. The capacity set is the quantity Q and D is the (unknown, random) annual demand for that model. The overage cost includes the savings from providing one unit less capacity, and the underage cost includes the lost profit from not being able to make one more car of that type. In reality, there are other considerations that must be included in making a capacity decision like this one, such as the option to operate the plant at one, two, or three shifts a day, and/or weekends. Plants also often are designed so that multiple models can be manufactured on the same assembly line, which provides the flexibility to make more of the hot selling models and fewer of the duds. It may also be possible to make adjustments to the capacity level during the year. All these considerations can be addressed in a model, and a newsvendor model can be a good starting point for the analysis.

## *Airline Overbooking*

Another application of the newsvendor model is to airline overbooking of flights. Tyler has another uncle, Howie, who works for an airline that faces, for each flight, the possibility that not all potential passengers who made reservations will show up for the flight. If one of Howie's flights has 100 economy seats, all 100 are reserved, but only 90 show up to board the flight, then the 10 empty seats generate no revenue, but would have if Howie had allowed 10 extra reservations to be made for the flight and all of them had shown up. The available quantity 100 of seats is fixed in advance by the type of airplane assigned to the flight. The decision $Q$ is the number of overbooked seats to allow. The (random) demand $D(Q)$ is the number of no-shows, passengers with reservations who do not show up for the flight, as a function of the number $Q$ of overbooked seats. If $Q > D(Q)$, then the flight is overbooked at the time of boarding: The number of no-shows is less than the number of overbooked seats, so there are some passengers who cannot board the flight. What the airlines do in this situation is offer vouchers redeemable on future flights to confirmed passengers who agree to take a later flight. Their initial offer is sometimes a free domestic flight within their system. Passengers who accept the offer are assumed to be appeased, creating no unforeseen negative consequences. If not enough passengers accept the offer, the airline will raise the offer, as in an auction, essentially paying the lowest price it can to reduce the number of passengers on the flight to its capacity.

The simplest version of this problem occurs when $D(Q)$ is independent of $Q$, which means that the distribution of the number of no-shows is not affected by the number of overbooked seats. In the next paragraph, we will briefly discuss the more realistic scenario of there being a dependence of $D$ on $Q$. The situation of overbooked confirmed passengers corresponds to $Q > D$, which, from the perspective of the newsvendor model, is an "overage" because the quantity set (number of overbooked seats) is higher than (over) the demand (number of no-shows). Thus, the unit "overage" cost here is the expected appeasement cost that the airline would save if it had overbooked one fewer seat and therefore had one fewer overbooked passenger at boarding. (It consists of the expected cost of adding one reservation to the later flight plus the expected cost of the voucher.) The unit "underage" cost is the expected additional return the airline would obtain by increasing the number of overbooked seats by one and therefore had one more paying customer on the flight. Both of these costs can be quite large and similar in magnitude. If they were equal, the critical fractile would be 0.5, which means the airline should run an equal risk of having empty seats on the flight and having overbooked passengers at boarding. If passengers are simply appeased with a voucher for a free flight of the same type in the future (along with the seat on a later flight), and the later flight is underbooked, then the cost of appeasing passengers would be less than the additional revenue from a paying passenger, because the cost of the seat on the later flight is insignificant and the voucher has the appeased passenger simply flying for free on another flight, which might not be overbooked. Therefore, the expected cost of the voucher is less than a paid fare. In this case, the critical fractile, which is the optimal probability of having overbooked passengers at boarding, is greater than 0.5.

If we assume more realistically that the number of no-shows does indeed depend on the number of overbooked seats (*D* depends on *Q*), then, provided the relationship is not perverse, marginal analysis still yields the optimal number of seats to overbook. The expected gain from overbooking one more seat is found as follows. If that ticket holder shows up and there is room on the flight (at boarding), then the airline will receive the additional revenue from the ticket. If that ticket holder shows up and there is no room on the flight, then the gain will equal the additional revenue received from the ticket and the loss will equal the opportunity cost of the seat that will be taken on the later flight plus the opportunity cost of the voucher that will appease the customer. There is nothing to stop the airline from overbooking as many seats as possible until these costs exceed the gain from the ticket sold. The financial consequences of the ticket holder not showing up must also be worked into the equation. For example, if the ticket is fully refundable, then the airline loses the revenue from the ticket and incurs the costs of processing both the initial ticket purchase and the refund. Airlines often offer much cheaper fares for non-refundable tickets that involve a substantial penalty to change to another flight. This arrangement presumably reduces the incidence of no-shows.

The overbooking tradeoffs discussed here also are relevant to hotels, restaurants, and other services that take reservations.

## *Flexible Medical Savings Accounts*

This application of the newsvendor model is to personal finance. Some US employers allow their employees to specify an amount to be taken out of their pay check, before taxes, that can be used to cover medical expenses that are not covered by the employee’s medical (and dental) insurance. At the urging of Shenn, Tyler is considering signing up to have $10 per month from his pay allocated to this account. He would therefore select $120 as the amount in his account for the year. He has good insurance coverage, but he faces a number of potential medical expenses that are not covered by any of his insurance plans. For example, he must co-pay for each office visit and a percentage of any medical procedure required. He also has no coverage for prescription eye glasses or for certain dental procedures. He is not sure what the total of these uncovered expenses will be for the year, but, with the help of Shenn, he has estimated that the probability distribution for the total is normal with a mean of $200 and a standard deviation of $20. If his uncovered expenses are more than $120, then he will get the full $120 back. Tyler’s marginal income tax rate is 25%, so, because the $120 was not taxed, Tyler has saved $30: If Tyler had not signed up for the program, he would have had to pay that $120 out of his after-tax earnings. So he would have had to earn \$160 in pretax earnings to yield the \$120 needed to pay the uncovered expenses. Under the program, he only needed to use \$120 of his pre-tax earnings to cover these expenses, so he can keep \$40 more in pretax earnings, which converts into \$30 in after-tax earnings. The downside of the program is that if Tyler's uncovered expenses are less than the \$120 he put aside, then he loses the unused funds.

This is a newsvendor model: $Q$ is the amount put into the program for the year, and $D$ is the total amount of uncovered expenses. In case of an overage, $Q > D$, then if he had reduced $Q$ by one dollar, he could have kept that pretax dollar, which converts into \$0.75 in after-tax dollars. So the unit overage cost is $c_O = 0.75$. In case of an underage, $Q < D$, then if he had increased $Q$ by one dollar, he would have lost the \$0.75 in after-tax dollars, but would have saved \$1 in after-tax dollars that he needed to use to pay the extra uncovered expenses. Thus, the (net) unit underage cost is $c_U = 0.25$.

The critical fractile, the optimal probability of having too much in the account, is therefore

$$\gamma = \frac{c_U}{c_U + c_O} = \frac{0.25}{0.25 + 0.75} = 0.25,$$

which, using Exhibit 7.5, means that Tyler should put

$$Q^* = \mu + z^* (\gamma)\, \sigma = 200 - 0.674(20) = 187$$

into his account.

## Other Applications

There are numerous other applications of the newsvendor model, such as setting cash reserves by a bank or an individual for a checking account, selecting spare parts for a product at the end of the life of its production line, deciding how much scrubber capacity to have available to clean noxious fumes being produced, college admissions, water reservoir management, staff sizing in a service business, and prepurchased maintenance calls. See Wagner (1975), Shogan (1988), and Denardo (2002) for more.

## Historical Background

The first generally accepted framing and discussion of what we now call the newsvendor problem appeared in Edgeworth (1888) in the context of a bank setting the level of its cash reserves to cover demands from its customers. His focus was on use of the normal distribution to estimate the probability that a given level is large enough to cover the total demand.

Morse and Kimball (1951) coin the term “newsboy” in introducing this problem, use marginal analysis, and implicitly provide the critical fractile solution. Arrow et al. (1951) formulate a more general problem and also implicitly provide the critical fractile solution within a special case of their solution. Whitin (1953) explicitly provides the critical fractile solution.

The problem began to be known as the *Christmas tree problem* and the *newsboy problem* in the 1960s and 1970s. In the 1980s, researchers in the field sought vocabulary, such as the “newsperson” model, that would not be gender specific. Matt Sobel’s suggestion, the *newsvendor problem*, is now in general use.

So far, all discussion has been limited to the case of proportional costs. When Arrow et al. (1951) quoted the defeated Richard III as exclaiming, “A horse, a horse, my kingdom for a horse,” they not only suggested that the unit underage (shortage) cost might be very high, but that it might get larger as a function of the size of the shortage. Early work, such as Dvoretzky et al. (1952) and Karlin (1958), included such nonproportional (nonlinear) costs in their analyses. Karlin (1958) also allowed for the possibility that the overage cost might be a nonlinear function of the overage.

Suppose that both the underage and overage cost functions are *convex*, which can be understood as being bowl shaped with the property that no matter how they are tilted, while being kept right side up, their only flat spots can be at the bottom. (The proportional cost case is covered.) Then the expected opportunity costs, as a function of the quantity selected, is convex, so any flat spot we find on it will be at the bottom and therefore minimize the expected costs. (Flat spots are found using calculus by computing the derivative and setting it to zero.) From a computational perspective, a solution is easily found, although not nearly as easily as computing a critical fractile $\gamma$ based on the given unit costs and then finding the $Q$ such that $P(D \leq Q) = \gamma$.

Bellman et al. (1955) provide the explicit critical fractile solution for the multiple replenishment problem of a nonperishable product in the stationary, infinite horizon case. Veinott (1965) and Sobel (1981) extend the analysis to the nonstationary and finite horizon cases and to other settings.

*Gallego (1997)* provides a realistic formulation of the salvage value of unsold units: He assumes that there is a random quantity that can be sold at a given unit salvage value. It is also possible to include possibly many different markets in which the unsold units can be salvaged. Each market is characterized by the unit salvage value that it offers and the (possibly random) maximum number that can be salvaged at that price. This approach avoids the surprise of having many more unsold units to salvage than were anticipated.

When the purchase costs for the resource are proportional to the amount purchased, then it is straightforward to reformulate the costs as proportional to the overage and the underage. However, in many realistic situations, the purchase costs are not proportional. For example, there may be economies of scale, in that the average purchase cost may decrease in the size of the purchase. Considerations such as this have led to the development of stochastic inventory theory, with the seminal contribution of Arrow et al. (1958). Zipkin (2000) and Porteus (2002) give expositions of the current status of that theory, with Harrison et al. (2003) giving an introduction to supply chain management, a much broader field that developed rapidly with the help of researchers in inventory management. There has been an analogous development of capacity management, with Manne (1961) providing an early impetus, and Van Mieghem (2003) giving a recent review. The development of revenue management is more recent, with Talluri and van Ryzin (2004) providing an exposition of its recent status. In short, the newsvendor model is a key component of several different modern disciplines that address practical problems.

**Acknowledgments** The author wishes to thank Matthew Sobel for his comments and suggestions leading to a much improved document and Linda Bethel for helping convert the author's LaTex document into Word.

## References

Arrow, K., T. Harris, J. Marschak (1951) "Optimal Inventory Policy," *Econometrica* **19** 250–272.

Arrow, K., S. Karlin, H. Scarf (1958) *Studies in the Mathematical Theory of Inventory and Production*, Stanford University Press, Stanford.

Bellman, R., I. Glicksberg, O. Gross (1955) "On the Optimal Inventory Equation," *Management Science* **2** 83–104.

Denardo, E. (2002) *The Science of Decision Making: A Problem-Based Approach Using Excel*, John Wiley, New York.

Dvoretzky, A., J. Kiefer, J. Wolfowitz (1952) "The Inventory Problem: I. Case of Known Distributions of Demand," *Econometrica* **20** 187–222.

Edgeworth, F. (1888) "The Mathematical Theory of Banking," *Journal of Royal Statististical Society* 113–127.

Gallego, G. (1997) "Uncertain Demand at Salvage Value," *Notes for IEOR 4000: Production Management*, Columbia University, New York.

Harrison, T., H. Lee, J. Neale 2003 *The Practice of SupplyChain Management: Where Theory and Application Converge*, Kluwer Academic Publishers.

Karlin, S. (1958) "One Stage Inventory Models with Uncertainty," in Arrow, K., S. Karlin, H. Scarf (eds) *Studies in the Mathematical Theory of Inventory and Production*, Stanford University Press, Stanford, 109–134.

Manne, A. (1961) "Capacity Expansion and Probabilistic Growth," *Econometrica* **29** 632–649.

Morse, P., G. Kimball (1951) *Methods of Operations Research*, Technology Press of MIT, Cambridge.

Porteus, E. (2002) *Foundations of Stochastic Inventory Theory*, Stanford University Press, Stanford.

Shogan, A. (1988) *Management Science*, Prentice-Hall, Englewood Cliffs.

Sobel, M. (1981) "Myopic Solutions of Markovian Decision Processes and Stochastic Games," *Operations Research* **26** 995–1009.

Talluri, K., G. van Ryzin (2004) *The Theory and Practice of Revenue Management*, Kluwer Academic Publishers.

Van Mieghem, J. (2003) "Capacity Management, Investment, and Hedging: Review and Recent Developments," *Manufacturing & Service Operations Management* **5** 269–302.

Veinott, A., Jr. (1965) "Optimal Policy for a Multi-product, Dynamic, Nonstationary, Inventory Problem," *Management Science* **12** 206–222.

Wagner, H. (1975) *Principles of Operations Research*, Prentice-Hall, Englewood Cliffs.

Whitin, S. (1953) *The Theory of Inventory Management*, Princeton University Press, Princeton.

Zipkin, P. (2000) *Foundations of Inventory Management*, McGraw-Hill/Irwin, New York.

# Chapter 8
# The Economic Order-Quantity (EOQ) Model

**Leroy B. Schwarz**
**Purdue University**

*The economic order-quantity model considers the tradeoff between ordering cost and storage cost in choosing the quantity to use in replenishing item inventories. A larger order-quantity reduces ordering frequency, and, hence ordering cost/month, but requires holding a larger average inventory, which increases storage (holding) cost/month. On the other hand, a smaller order-quantity reduces average inventory but requires more frequent ordering and higher ordering cost/month. The cost-minimizing order-quantity is called the Economic Order Quantity (EOQ). This chapter builds intuition about the robustness of EOQ, which makes the model useful for management decision-making even if its inputs (parameters) are only known to be within a range of possible values. This chapter also provides intuition about choosing an inventory-management system, not just an EOQ.*

## Introduction

Lauren Worth is excited about her new job at Cardinal Hospital. Lauren had worked at Cardinal, first as a candy-striper, and then, after college, as a registered nurse. After several years "working the wards," Lauren left Cardinal to get an MBA. Now, having recently graduated, Lauren had returned to Cardinal as its first "Inventory Manager."

Lauren's boss, Lee Atwood, is Purchasing Manager at Cardinal Hospital. Lauren has known Lee since her days as a candy-striper, and regards him as a friend. Lee describes himself as being from the "old school" of inventory management. "Lauren, I know that there are sophisticated methods for managing inventories, and, in particular for determining 'optimal order-quantities.' And, I know that Cardinal is paying a 'management-cost penalty' for using other than optimal order-quantities. In other words, that Cardinal is incurring higher than the minimum possible inventory-management costs in managing its inventories. But, frankly, I'm skeptical about what it will cost Cardinal Hospital to reduce that management-cost penalty. That's because determining the optimal order-quantity will require two things that are in short supply around here."

D. Chhajed and T.J. Lowe (eds.) *Building Intuition*: *Insights From Basic Operations Management Models and Principles*.
doi:10.1007/978-0-387-73699-0, © Springer Science + Business Media, LLC 2008

"First, is the knowledge about costs and demands. For example, in the general supplies area, where I would like you to begin, in order to use the 'EOQ model,' you need to know what it costs to place an order, and how much it costs to hold an item in inventory. Now, I know that it costs Cardinal money to place an order and money to hold an item in inventory, but we don't know what these costs are. Given enough time and trouble, we can find out what those numbers are, but, at best, they would be estimates. Now, let's suppose it costs, say, $100 a year to get those estimates for each of the 2,000 general-supply items we are talking about, and to keep those numbers up to date. Once we plug those numbers into this formula and determine the 'optimal' order-quantity, will using that order-quantity save us at least the $100/year that it is costing us to feed that formula? In other words, if we save $90/year by making better decisions, but it costs us $100/year to make those better decisions, then we have lost $10/year!"

"Another thing that's been in short supply, at least until you've arrived, is enough expertise to know which inventory-management system we should use to manage the different inventories we manage. For example, as you know, some of these items are expensive, others are very inexpensive; some perish, some don't; some become obsolete quickly, others don't. I know there are lots of different systems out there—EOQ, JIT, MRP, and the 'newsvendor,' to name just a few—but each of these has overhead connected with it. It is logical that the more sophisticated the model, the 'better' the order-quantity decision, but it also makes common sense that the more sophisticated model will be more expensive to install and keep up to date. For example, suppose that the EOQ model would reduce ordering and holding costs by $200/year, but cost us only $100/year to use and keep up to date: a net saving of $100/year. Would some other model—JIT, for example—net Cardinal a larger saving?"

Lauren was confused. She had thought that all she had to do was to plug numbers into the EOQ model and Cardinal would be making optimal order-quantity decisions. She had not thought about the cost of getting those numbers; or, worse, the fact that those numbers would only be estimates that would have to be updated over time. As a consequence, it had not occurred to her that Lee Atwood's policy of ordering a 3-month supply for *everything* might, all things considered, be a good decision-rule. Finally, Lauren also had not considered the possibility of using a more sophisticated system than EOQ, one that might yield larger savings but have a higher overhead cost.

## The Decisions to be Made

One of the most frequent decisions faced by operations managers is "how much" or "how many" of something to make or buy in order to satisfy external (e.g., customer demand) or internal requirements for some item.

Many times, this decision is made with little or no thought about its cost consequences. Examples are "order a case" or "make a month's supply." Lee Atwood's policy is to order a 3-month supply whenever he ordered anything. In *some* business scenarios, such decision-rules might be OK, in the sense that they require very little information (e.g., what *is* a month's supply?), are easy to implement, and yield inventory-management costs that are only a few percent above the minimum possible cost (i.e., a small management-cost penalty). Indeed, such "thoughtless" decisions might even be optimal, in the sense that they might minimize the *total* cost of inventory management, including the cost of the information and implementation requirements of the decision-rule being used. However, in other scenarios, decision-rules like these can be quite bad; in other words, their associated management cost-penalty is very large and, an alternative decision-rule, one that required more expensive information and/or implementation would yield a considerably smaller total cost of inventory management.

So, the *fundamental* decision about the "order-quantity decision" *isn't* "What's the right quantity"? Instead, it is: "Under what circumstances does the *choice* of the order-quantity make a lot of difference (in terms of the management-cost penalty)"? After all, if the choice of the order-quantity does not make a significant difference, then common sense suggests that the choice of the order-quantity does not matter much; and, therefore, should be made with a focus on making and implementing the order-quantity decision as inexpensively as possible (in terms of its information and implementation costs). For example, at Cardinal Hospital, Lee Atwood's 3-month supply policy might be fine; that is, using EOQ—or some even more sophisticated system—would cost more than it was worth. On the other hand, if the choice of the order-quantity *does* make a significant difference, then the order-quantity decision may deserve a great deal of management attention and thoughtful judgment. Or, to quote Ford Harris (1913), who created the EOQ model, "This is a matter that calls, in each case, for a trained judgment, for which there is no substitute."

This chapter will address the question of whether or not the choice of the order-quantity makes a lot of difference in terms of its management-cost penalty, using the "Economic Order Quantity (EOQ) model."

Assuming that the order-quantity decision *does* make a significant difference, the question *then* becomes "What's the right quantity"? As might be expected, the larger the management-cost penalty, the more difficult it is to answer this question, typically because of costs and/or constraints that arise from the complexity of each specific business scenario. In such scenarios, the insights provided by the analysis of the EOQ model (see below) and the so-called "EOQ formula" are often helpful in guiding management about the order-quantity decision. Or, to quote Ford Harris again, in such scenarios, "… using the formula as a check, is at least warranted."

## The EOQ Business Scenario

The EOQ business scenario is simple. It involves selecting the order-quantity "Q" that minimizes the average inventory-management cost/time—where "time" can be a "year," a "month," or a "week"— for an item with demand that goes on forever at a demand *rate* that never changes. Pretty simple, right? So simple, that if the value of the insights it provides depended on being able to find even *one* real-world scenario that matched it, then it would be worthless. But, therein lies the tale…..

## *Customer Demand*

The customer demand (or internal usage, as at Cardinal Hospital) rate of the item is assumed to be known and constant at a rate D units/time. The “units” dimension might be cases or gallons or truckloads; it does not matter. For convenience we will measure D in units of one and time in years.

## *Leadtime*

The EOQ business scenario also assumes that the order (or manufacturing) leadtime (i.e., the time interval between placing the order and receiving the corresponding order quantity) is zero. In other words, that delivery or manufacturing is instantaneous. This assumption (conveniently) removes the question of “When to order?” the chosen order-quantity: Order Q each time inventory falls to zero.

Although the instantaneous-delivery assumption is obviously unrealistic—after all, even in Star Trek movies the Replicator takes a few moments to produce what Captain Picard orders—any *fixed* leadtime is easily accommodated, provided that demand is known and constant: Order Q when inventory equals a leadtime’s supply. For example, if the leadtime is one week, then order Q when inventory equals a week’s supply.

On the other hand, if leadtimes are unpredictable, then even given a known and constant demand rate, D, the question “When to order?” becomes much more difficult to answer. We will address this question briefly below.

## *Costs*

The EOQ business scenario accounts for three types of costs: (1) the cost of the units themselves; (2) the cost of holding units in inventory; and (3) the fixed order (or manufacturing set-up) cost.

*Unit Cost:* The cost of the units themselves, denoted *c*, and measured in $/unit, is assumed to be fixed, regardless of the number of units ordered or manufactured; although quantity discounts or other economies of scale are ignored in the *basic* EOQ model, but are easily accommodated (See below).

*Inventory-Holding Cost:* The inventory-holding cost in the EOQ business scenario, denoted *h*, represents management’s cost of capital (i.e., the time value of money) invested in the units, the cost of the space consumed by the units, taxes or insurance premiums on the inventory itself, and, in some cases, allowances for obsolescence (i.e., the possibility that the items might becomes useless while they are sitting in inventory) or “shrinkage” (i.e., that units may be lost or stolen).

The inventory-holding cost, *h*, is measured in \$/(unit × time), where the “unit” and the “time” dimensions are the same as used in defining the demand rate D. So, for example, if the demand rate is measured in gallons/year, then *h* is measured in \$/(gallon × year); in other words, the dollar cost to hold one gallon for one year’s time. For convenience, as above, we will measure time in years.

Every thoughtful inventory manager recognizes that holding inventory costs money. In many scenarios, the value of *h* is likely to be between 25% and 50% of the cost of the item itself, depending on the company’s cost of capital—for example, what it costs the company to borrow money—and the item’s risk of obsolescence. On the other hand, it is typically very difficult to determine *h*’s exact dollar value. For example, does it cost, say, \$1 to hold a unit of an item in inventory for a year; or does it cost \$2? Most company’s cost-accounting systems don’t report this number, although every accountant knows that it’s real. It’s even possible that, over time, the value of *h* changes. For example, if storage space is ample at some times of the year and tight at others, or if a company’s cost of capital changes over time, then the value of *h* changes over time.

For the time being, we will assume that the value of *h* is known and constant, and examine the importance of this assumption below.

*Fixed Order Cost:* The fixed order cost in the EOQ business scenario, denoted *K*, and measured in dollars, represents all the costs associated with placing an order *excluding* the cost of the units themselves. In other words, *K*, represents any administrative (i.e., paperwork) costs of placing and/or receiving an order (e.g., inspection); and, in a manufacturing scenario, the cost of setting up the equipment to produce the order-quantity.

Like inventory-holding costs, fixed order costs are real costs for every company; and, in some scenarios, very significant costs. A set-up (i.e., equipment changeover) in pharmaceutical manufacturing, for example can shut down production and engage dozens of workers for weeks at a time; thereby costing thousands of dollars. However, like inventory-holding cost, the challenge is determining its exact dollar value. For example, does it cost, say, \$100 or \$300 to place an order? Most cost-accounting systems do not report this number, either. It is even possible that, like the inventory-holding cost *h*, the value of *K* changes over time. For example, if the ordering, set-up, and/or receiving departments are busy at some times of the year but idle at other times, then the value of *K* itself changes over time.

For the time being, we will assume that the value of *K* is known and constant, and examine the importance of this assumption below.

*Picking the “Optimal” Order-Quantity:* In the EOQ business scenario, the goal is to select the order-quantity, Q, that minimizes the average inventory-management cost/time—measured in \$/year—that results from ordering Q (whenever inventory reaches zero), and doing that (repetitively) forever (i.e., over an “infinite time horizon”). We will denote this ***a***verage ***c***ost resulting from ***Q*** as “AC(Q)”

Given the three costs accounted for in the EOQ scenario it follows that AC(Q) must be:

$$AC(Q) = K \times \frac{D}{Q} + h \times \frac{Q}{2} + c \times D. \tag{1}$$

That is, as described above, the average cost/year of using *Q* as the order-quantity has three parts, the average annual cost of ordering (first term), the average annual inventory-holding cost (second term), and the average annual cost of the units themselves.

Taking a closer look, the average annual ordering cost, *K* × (*D*/*Q*), equals the fixed order cost *K*, multiplied by the average number of orders placed per year, *D*/*Q*. (For example, if the demand rate *D* is 1,200 units/year and *Q* = 200 units, then, on average *D*/*Q* = 1,200/200 = 6 orders will be placed each year.) The middle term in (1), *h* × (*Q*/2), is the inventory-holding cost rate, *h*, multiplied by the average inventory, *Q*/2. (The average inventory is *Q*/2 since the maximum is *Q*, the minimum is zero, and it falls from *Q* to zero at a uniform rate.) Finally, the average annual cost of the units themselves is *c* × *D*; that is, the cost/unit of the item being inventoried, multiplied by its annual demand.

The first thing to notice about AC(Q) is that the average annual cost of the units themselves *does not depend* on Q; that is, whether we buy in small quantities (e.g., place several orders per year) or in large quantities (e.g., order every other year), the average cost of the units themselves doesn't depend on the order-quantity *Q*. Consequently, this term, *c* × *D*, is typically left out of *AC*(*Q*), as in (2) below:

$$AC(Q) = K \times \frac{D}{Q} + h \times \frac{Q}{2}. \tag{2}$$

Written in this manner, *AC*(*Q*) is the sum of two terms, one of them: *K* × (*D*/*Q*), which decreases as *Q* increases; and the other, *h* × (*Q*/2), which increases as *Q* increases. In managerial terms, this makes perfect sense: The average annual ordering cost increases as *Q* decreases because more orders must be placed each year; and the average annual inventory-holding cost increases as *Q* increases because the average inventory, *Q*/2, increases. Figure 8.1 provides a graph of *AC*(*Q*) and its two components: *K* × (*D*/*Q*) and *h* × (*Q*/2).

Note that the average ordering cost, *K* × (*D*/*Q*), decreases as *Q* increases. For example, if *D* = 1,200 units/year, then as *Q* increases from, say, 100–1,200, then the average number of orders/year, *D*/*Q*, decreases from 12 to 1 time/year. (Notice also, that if, say, *Q* = 2,400, then one order will be placed every 2 years, and *D*/*Q* = ½.) So, if management wanted to minimize the average order cost, then it would choose the largest possible value of *Q*. For example, this might be the largest *Q* that it could afford to buy or the most it could store in the space available. On the other hand, the average inventory-holding cost, *h* × (*Q*/2), increases as *Q* increases. In other words, the larger the order-quantity, the larger the average inventory, and,

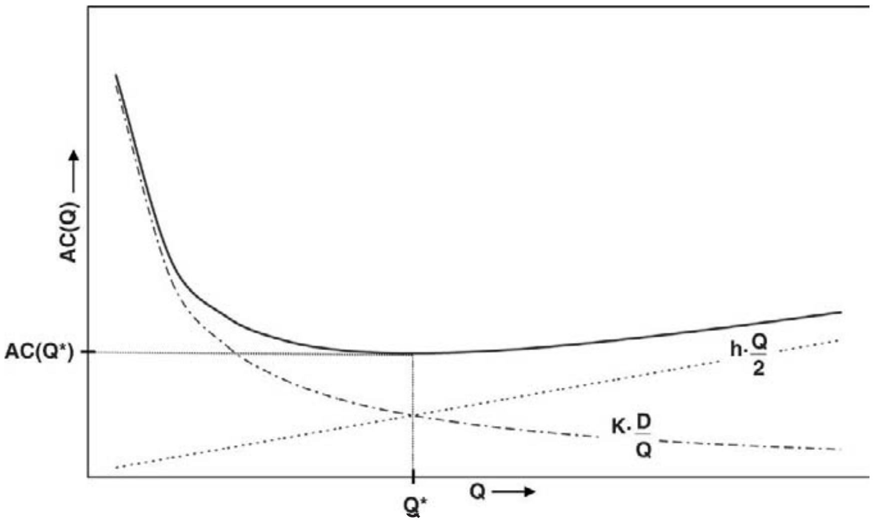


**Fig. 8.1** A Graph of *AC*(*Q*) Versus *Q*

hence, the larger the average inventory-holding cost. So, if management wanted to minimize the average inventory-holding cost, then it would order the smallest possible value of *Q*. For example, this might be the minimum that the supplier would sell or the smallest quantity that could be made.

Hence, the choice of *Q* is a trade-off between the average ordering cost and the average inventory-holding cost. Note in Fig. 8.1 that the minimum value of *AC*(*Q*) occurs where the average annual ordering costs and average annual inventory-holding graphs cross; that is, where $K \times D/Q) = h \times (Q/2)$. Thus, we can find the optimal *Q*, denoted *Q** by solving equation (3) for *Q**; that is, solving

$$K \times \frac{D}{Q*} = h \times \frac{Q*}{2} \tag{3}$$

for *Q**. The solution is:

$$Q* = \sqrt{\frac{2 \times K \times D}{h}}. \tag{4}$$

This is the so-called “square-root” or EOQ formula for *Q**.

To determine *AC*(*Q**), the inventory-management cost/year using $Q = Q*$, one substitutes (4) into (2). After some algebra:

$$AC(Q*) = \sqrt{2 \times K \times D \times h}. \tag{5}$$

## A Closer Look at the Assumptions Of EOQ: Sensitivity Analysis

In order to develop the EOQ model, several unrealistic assumptions must be made. The most *unrealistic* of these is that the demand rate is known and constant. However, even if management were comfortable with that assumption, there is the practical problem of estimating the values of the fixed order cost (K) and the inventory-holding cost ($h$) to "plug" into equation (4).

In order to assess the impact of all these assumptions, we are going to do "sensitivity analysis." That is, we are going to assess the consequences of these assumptions on the costs we experience because we are making assumptions that might not be—indeed, probably are not—valid. We are going to conduct this sensitivity analysis in several steps. In Step 1 we will assume that all of the assumptions of the EOQ business scenario are valid, but that we ignore the EOQ formula, and, instead, select $Q$ without much thought. In Steps 2 and 3, we will continue to assume that EOQ model's assumptions hold, and that we use the EOQ formula, but that we only have very rough estimates about its inputs: $K$, $D$, and $h$. In Step 4 we will use the insight provided in Steps 2 and 3 to examine the model's assumptions about known and constant demand and known and constant cost parameters. We will continue to assume that production and/or delivery are instantaneous, in order to focus on the order-quantity decision.

### *Sensitivity Analysis: Step 1*

We begin with a business scenario in which *all* of the assumptions of the EOQ business scenario are valid. We even know the true values of $D$, $h$, and $K$; *but* we select a value of $Q$ without much thought (e.g., order a case). We will denote our choice of Q as $\hat{Q}$. It is unlikely, of course, that our chosen $\hat{Q}$ will equal $Q^*$. Since there is only one $Q^*$ that minimizes $AC(Q)$, we will be paying a management-cost penalty for using $\hat{Q}$ instead of $Q^*$; that is $AC(\hat{Q}) \geq AC(Q^*)$, and the difference, $[AC(\hat{Q}) - AC(Q^*)]$, is the penalty. For example, if our $\hat{Q}$ is larger than $Q^*$, then, in Fig. 8.1, $\hat{Q}$ will be to the right of $Q^*$ and we will experience $AC(\hat{Q})$ larger than $AC(Q^*)$. How much higher, of course, depends on how much larger $\hat{Q}$ is than $Q^*$ and on the shape of the $AC(Q)$ function. Suppose, for example that our chosen $\hat{Q}$ is twice the size of $Q^*$; i.e., $\hat{Q} = 2 \times Q^*$. The question we want to answer is: *Will the management-cost we pay, $AC(\hat{Q})$, be 2 times $AC(Q^*)$? Will it be more, or will it be less?*

Specifically, we are interested in the ratio $AC(\hat{Q})/AC(Q^*)$. Notice that if, by chance, $\hat{Q} = Q^*$, then, of course, $AC(\hat{Q}) = AC(Q^*)$, and the "penalty-cost ratio" would equal 1. Otherwise, this ratio would be above 1, say, 1.25, which indicates that the management penalty cost is 25% above the minimum possible $AC(Q^*)$. Using equation (2), once with $Q = \hat{Q}$ and once with $Q = Q^*$ we get:

$$\frac{AC\left(\breve{Q}\right)}{AC\left(Q*\right)} = \frac{K \times \frac{D}{\breve{Q}} + h \times \frac{\breve{Q}}{2}}{K \times \frac{D}{Q*} + h \times \frac{Q*}{2}}. \tag{6}$$

Using the fact that $K \times (D / Q^*) = h \times (Q*/2) = \frac{\sqrt{2 \times \mathrm{K} \times \mathrm{D} \times \mathrm{h}}}{2}$, after a little algebra,

$$\frac{AC\left(\breve{Q}\right)}{AC\left(Q*\right)} = 0.5 \times \left[\frac{Q*}{\breve{Q}} + \frac{\breve{Q}}{Q*}\right]. \tag{7}$$

What equation (7) says is that the ratio of the management cost we *will* experience—from using the "wrong" $Q$, $\hat{Q}$— to the minimum possible management cost we *would* experience—if we used the optimal $Q$, $Q^*$—is one-half the sum of Q*/$\hat{Q}$and $\hat{Q}$/Q*. So, for example, if our chosen $\hat{\mathrm{Q}}$= 2× Q* then

$$\frac{AC\left(\breve{Q}\right)}{AC\left(Q*\right)} = 0.5 \times [0.5 + 2] = 1.25.$$

In other words, the average annual cost we will experience will be 25% higher than the minimum possible. Note that a 100% error in choosing $\hat{Q}$has yielded a management-cost penalty of only 25% of $AC(Q^*)$.

Table 8.1 provides the values of $AC(\hat{\mathrm{Q}})/AC(Q^*)$ for $\hat{Q}$ranging from 0.1 $Q^*$ to 10$Q^*$. Note, that the penalty-cost ratio is symmetric around 1 (e.g., that the ratio is the same for $\hat{Q}/Q^* = 0.1$ as for $\hat{Q}/Q^* = 10$).

The management intuition from Step 1 is that as long as $\hat{Q}$ is in the "ballpark" of $Q^*$ (e.g., within a factor of 2), then the penalty-cost ratio is relatively small (e.g., less than or equal to 1.25), whereas if $\hat{Q}$ is "way off" (e.g., more than 10 times larger or smaller), then the penalty-cost ratio will be quite large (e.g., greater than or equal to 5.050). Although this is a *valuable intuitive insight*, its *practical value* is limited by the assumption that we know what $Q^*$ is. This leads to Step 2:

## *Sensitivity Analysis: Step 2*

Consider a closely related business scenario in which: (a) all of the assumptions of the EOQ business scenario are valid; (b) management has chosen to use the EOQ formula to compute the order quantity; (c) management *knows* the exact dollar values of $K$, the fixed order cost (e.g., $K$ = \$100), and $h$, the inventory-holding cost (e.g., $h$ = \$2/unit × year); but (d) *doesn't know* what the demand rate, $D$, is. Instead, management only knows the range of possible values for $D$; for example, that $900 \leq D \leq 1{,}600$; but doesn't know for certain what the value of $D$ is.

**Table 8.1** $AC(\hat{Q})/AC(Q^*)$ Versus $\hat{Q}/Q^*$

| $\hat{Q}/Q^*$ | $AC(\hat{Q})/AC(Q^*)$ |
|---|---|
| 0.1 | 5.050 |
| 0.15 | 3.408 |
| 0.2 | 2.600 |
| 0.25 | 2.125 |
| 0.3 | 1.817 |
| 0.35 | 1.604 |
| 0.4 | 1.450 |
| 0.45 | 1.336 |
| 0.5 | 1.250 |
| 0.55 | 1.184 |
| 0.6 | 1.133 |
| 0.65 | 1.094 |
| 0.7 | 1.064 |
| 0.75 | 1.042 |
| 0.8 | 1.025 |
| 0.85 | 1.013 |
| 0.9 | 1.006 |
| 0.95 | 1.001 |
| 1 | 1.000 |
| 1.05 | 1.001 |
| 1.1 | 1.005 |
| 1.25 | 1.025 |
| 1.5 | 1.083 |
| 2 | 1.250 |
| 5 | 2.600 |
| 6 | 3.083 |
| 7 | 3.571 |
| 8 | 4.063 |
| 9 | 4.556 |
| 10 | 5.050 |

Let's begin by assuming that the value of $D$ = 900 units/year. Using equation (4) yields

$$Q^* = \sqrt{\frac{2 \times K \times D}{h}} = \sqrt{\frac{2 \times 100 \times 900}{2}} = 300 \text{ units/order.}$$

Similarly, if we assume that the value of $D$ = 1,600 units/year, then

$$Q^* = \sqrt{\frac{2 \times K \times D}{h}} = \sqrt{\frac{2 \times 100 \times 1600}{2}} = 400 \text{ units/order.}$$

Based on these calculations we *have not learned* the value of $Q^*$, since we remain ignorant about the value of $D$. However, we *have learned* that $Q^*$ must be between 300 and 400 units/order!

As a consequence, we know that if we chose $\hat{Q}$= 300 (assuming $D$ = 900) when, in fact, $D$ = 1,600, and, hence, $Q^*$ = 400, then $\hat{Q}/Q^*$ = 300/400 = 0.75; and, using (7), the cost-penalty ratio would be:

$$\frac{AC\left(\breve{Q}\right)}{AC\left(Q^*\right)} = 0.5\left[\frac{Q^*}{\breve{Q}} + \frac{\breve{Q}}{Q^*}\right] = 0.5\times\left[0.75 + \frac{1}{0.75}\right] = 1.0417.$$

In other words, if $D$ was actually 1,600 units/year, and we made the worst possible mistake in estimating its value (i.e., that $D$ was 900 units/year), then the penalty-cost ratio would be 4.17% of $AC(Q^*)$. Since $AC(Q^* = 400) = \$800$/year, the management-cost penalty would be 0.0417 × \$800 = \$33.36/year.

Similarly, if we chose $\hat{Q}$= 400 (assuming $D = 1600$) when, in fact, $D = 900$ (and, hence, $Q^* = 300$), then $\hat{Q}/Q^* = 400/300 = 1.333$. The penalty-cost ratio would, again, be 1.0417, as before. However, in this case, $AC(Q^* = 300) = \$600$/year. Hence, the management-cost penalty would be 0.0417 × \$600 = \$25.02/year.

Based on these calculations, we *have not learned* the what $Q^*$ is, but we *have learned* that if we chose any estimate of $D$ in the range [900, 1,600], then the maximum penalty-cost we would experience would be 4.17% of the minimum possible cost $AC(Q^*)$. Of course, in choosing a value for $D$ to use in a scenario like this, management probably would not choose one of the extreme values (e.g., 900 or 1,600), but, instead, choose an estimate in between. For example, if management's estimated demand rate—we will denote this "$\hat{D}$" henceforth—was $\hat{D}$= 1,200 units/year, then the maximum penalty-cost ratio would be 1.010. In other words, the corresponding management-cost penalty would be just 1% above the minimum possible (of either \$800 or \$600/year).

## *Sensitivity Analysis: Step 3*

In the Step 3 we extend the assumptions of Step 2, except, now, management is not certain about the exact values for *any* of its inputs. So, in addition to choosing a value for $\hat{D}$, management also has to choose estimates for $K$, and $h$, which we will denote $\hat{K}$ and $\hat{h}$, respectively.

To take a specific example, suppose management knows only that:

$900 \le D \le 1{,}600$ unit per year,
$\$75 \le K \le \$125$/order, and
$\$1.50 \le h \le \$2.50$/unit × year.

As in Step 2, in order to compute the maximum management-cost penalty, we have to consider the most extreme estimates $(\hat{D},\hat{K},\hat{h})$ that management might choose and the most extreme true values that the $(D,K, h)$ might take on. In order to simplify the process, we can use the fact that the penalty-cost ratio depends only on the ratio $\hat{Q}/Q^*$; and, in particular, that, whatever the values of $\hat{Q}$ and $Q^*$, it is those values that generate either the smallest possible ratio $\hat{Q}/Q^*$or the largest possible ratio $\hat{Q}/Q^*$ that will yield the largest penalty-cost ratio.

Continuing, we can determine the largest possible $\hat{Q}/Q^*$ by asking: *what is the largest possible $\hat{Q}$ (based on its chosen $\hat{D}$,$\hat{K}$, $\hat{h}$,) that management might compute, using the EOQ formula; and what is the smallest that $Q^*$ might possibly be?*

Since EOQ is increasing in $D$ and $K$ and decreasing in $h$, the largest possible $\hat{Q}$ that management might choose is 516.40 (using $\hat{D}$= 1,600, $\hat{K}$= 125, and $\hat{h}$ = 1.50). Similarly, the smallest possible $Q^*$ is 232.39 (based on $D$ = 900, $K$ = 75, and $h$ = 2.50). Hence, the largest possible ratio $\hat{Q}$ /$Q^*$ ratio in this scenario equals 2.22 (=516.40/232.39). Using the same logic, the smallest possible ratio $\hat{Q}/Q^*$ ratio in this scenario equals 0.45 (1/2.22 = 232.39/516.40).

Plugging either of these $\hat{Q}$ /$Q^*$ ratios into equation (7) yields the largest possible penalty-cost ratio that management might experience, regardless of the value of $\hat{Q}$ it computes (that is, regardless of its estimates of $D$, $K$, and $h$; and regardless of what the true values of these parameters might be). This maximum ratio is

$$\frac{AC\left(\breve{Q}\right)}{AC\left(Q^*\right)} = 0.5 \times \left[2.22 + \frac{1}{2.22}\right] = 1.335.$$

In managerial terms, the analysis above tells us that in this EOQ business scenario, but with considerable uncertainty about the values of the parameters ($D$, $K$, $h$) to use in the EOQ formula, the *worst* management-cost penalty we would pay would equal 33.5% above the minimum possible cost (i.e., if we *knew* what the precise values of ($D$, $K$, $h$) were.

Finally, as in Step 2, management faced with this scenario would most likely choose in-between values for $\hat{D}$, $\hat{K}$, and $\hat{h}$, not their possible extremes. For example, if management chose to use $\hat{D}$= 1200, $\hat{K}$ = \$96.82, and $\hat{h}$ = \$1.94, then the maximum management-cost penalty would be 8.11% of AC(Q*).

The management intuition from Step 3 is that as long as the assumptions of the EOQ business scenario are valid, *one can use the EOQ formula, (4), to compute an order-quantity that results in a relatively small penalty-cost ratio even if management has a great deal of uncertainty about the values of D, K, and h to "plug" into it!*

## *Sensitivity Analysis: Step 4*

Finally, we relax the most significant of the EOQ model's assumptions: namely, that the demand rate, $D$, the fixed order cost, $K$, and the inventory-holding cost, $h$, are stationary (i.e., never changing).

Suppose, now, that the demand rate $D$ varies over some known range, [$\mathrm{D}_{\mathrm{Min}}$, $D_{\mathrm{Max}}$]. For example, suppose $D_{\mathrm{Min}}$ = 900 units/year and $D_{\mathrm{Max}}$ = 1600 units/year. Note that the difference between this scenario and the one above is that $D$ is no longer fixed at some unknown value. Instead, *D varies*, perhaps in an unknown manner, within this range. Similarly, suppose that both $K$ and $h$ vary, again, perhaps in some unknown manner, over their respective ranges; i.e., \$75 ≤ $K$ ≤ \$125/order and \$1.5 ≤ $h$ ≤ \$2.50/unit × year.

Common sense suggests that since $D$, $K$, and $h$ all change their respective values over time, then the optimal order-quantity policy probably should, too. In other words, the optimal policy probably orders different quantities at different times, depending on how predictable the pattern of changes in *D*, *K*, and *h* are and depending how much management knows about them. Instead, we will use a "stationary" order-quantity policy (i.e., one that orders the same quantity every time), and compute that order-quantity, $\hat{Q}$, using stationary (i.e., fixed) estimates of their true values.

For the sake of convenience, we will use the values $\hat{D}$ = 1,200, $\hat{K}$= $96.82, and $\hat{h}$ = $1.94, from Step 3. The corresponding $\hat{Q}$= 346.09. Recall from the analysis in Step 3, that regardless of whatever the true stationary values of *D*, *K*, and *h*, this order-quantity will never yield a management-cost penalty larger than 8.11% of *AC*(*Q**).

To see how this result applies when the values of *D*, *K*, and *h* are changing over time, imagine some moment of time when inventory reaches zero units. As described, management will order $\hat{Q}$= 346.09 units. In order to minimize the rate at which ordering and inventory-holding costs are accumulating *at that moment of time*, management should be ordering the *Q** that corresponds to whatever the true values of *D*, *K*, and *h* are *at that moment of time*. We do not know those values at that moment, but we do know the ranges for each. So, based on the analysis in Step 3, we can conclude that whatever their values are, in ordering $\hat{Q}$ we will never be accumulating a management-cost penalty of more than 8.11% above the minimum possible (associated with whatever *Q** is at that moment of time).

Exactly the same argument applies at *any* moment of time. That is, at any moment of time, inventory-management costs are accumulating at a rate based on management's chosen (stationary) EOQ policy. If, at the same moment of time, management had been using the optimal policy, inventory-management costs would be accumulating at the minimum possible rate. However, at *no* moment of time will management ever be accumulating an inventory-management-cost penalty at a rate higher than 8.11% of *AC*(*Q**).

See Lowe and Schwarz (1983) for a formal analysis of this management scenario and the conclusion stated above and interpreted below.

## The Bottom Line

The management intuition from the sensitivity analysis above is that in *any* EOQ-like management scenario—one with a demand (or usage) rate, a fixed order cost, and an inventory-holding cost—the EOQ formula will provide order-quantities that have relatively small management-cost penalty percentages, regardless of whether the demand rate is either known or constant and regardless of whether or not its associated ordering and inventory-holding costs are either known or constant. "Relatively small" means that the penalty-cost percentage is small (e.g., 8.11%) compared to the possible ranges of values for *D*, *K*, and *h*.

> For Lauren Worth, this means, first, that the EOQ model can be used to make "good" order-quantity decisions—that is, decisions with small management-cost penalty percentages—whether or not Cardinal Hospital's accounting system is able to yield

> accurate estimates of costs and/or demands. Furthermore, the EOQ model will make "good" order-quantity decisions even if the cost and demand inputs to the EOQ model are nonstationary (i.e., change their values over time).

This insight is of tremendous importance: It is what accounts for the use of the EOQ model and EOQ formula in so many real-world inventory-management applications.

## Back to the Fundamental Decision

We began this chapter by arguing that the *fundamental* decision about the "order-quantity decision" *is not* "What's the right quantity"?. Instead, it is: "Under what circumstances does the *choice* of the order-quantity make a lot of difference (in terms of the management-cost penalty)"?. And, we promised that we would address this question using the EOQ model. We are now ready to do so.

Consider managing the inventory of some item, or, practically speaking, a set of items with roughly the same ranges of values for their fixed order cost, inventory-holding cost, and demand rates, respectively. Apply the analysis above in order to determine the maximum inventory-management penalty-cost percentage, assuming that the item's order-quantity would be chosen using the EOQ formula and "ball-park" estimates of it inputs: $\hat{D}$, $\hat{K}$, and $\hat{h}$. Denote this percentage P%. To illustrate, in Step 4 above, we determined that the maximum management-cost penalty, P%, would be 8.11% of $AC(Q^*)$.

Now, the question is: "Is P% a large dollar amount or not?" For example, is an 8.11% management-cost penalty in Step 4 a significant amount of money or not? The answer, of course, depends on the true, but unknown values of $D$, $K$, and $h$, but we can put an upper limit on it by assuming that the true values of $D$, $K$, and $h$ are their maximum possible values: $D_{\text{Max}}$, $K_{\text{Max}}$, and $h_{\text{Max}}$.

In Step 4 these values are $D$ = 1,600, $K$ = \$125, and $h$ = \$2.50. Based on these, the corresponding $AC(Q^*) = \sqrt{2 \times 125 \times 1600 \times 2.50} = \$1{,}000/\text{year}$. Hence, the maximum possible management-cost penalty incurred by managing this item using the EOQ model is \$81.10/year = 0.0811 × 1,000.

Finally, management must ask itself: Would the information and implementation costs associated with managing this item using a more sophisticated model—one that took into account the fluctuating values of $D$, $K$, and $h$—cost more or less than \$81.10/year? Remember, the maximum possible reduction in annual ordering and inventory-holding costs resulting from using this more sophisticated model is \$81.10.

If the answer is that this more sophisticated model would cost more than \$81.10/year, then management should choose to use the EOQ model to select $\hat{Q}$. Why? Because even if this more sophisticated model were able to reduce inventory-management costs to the minimum, since it cost more than it was worth, it would actually increase the total management costs associated with this item.

If, on the other hand, the answer is that this more sophisticated model would cost less, then, management should test this model's performance against that of the EOQ model. Most likely this will require a computer simulation of the management scenario, one that would keep track of the total inventory-management costs of both models—EOQ and the more sophisticated model—including whatever the increased information and implementation costs of the more sophisticated model are.

In other words, by determining the maximum management-cost penalty/year that is associated with the use of the EOQ order-quantity, management can determine the maximum savings—that is, the maximum possible reduction in penalty cost—that might be derived from the use of a more sophisticated system.

In the same manner, management can estimate the maximum management-cost penalty associated with any order-quantity decision-rule, such as "Order a case." or "Order a month's supply." All that is required are estimates of the maximum and minimum possible values for $D$, $K$, and $h$, as in Steps 3 and 4 above. Then, determine the minimum and maximum values of $\hat{Q}/Q^*$ to determine the penalty-cost ratio $AC(\hat{Q})/AC(Q^*)$. Finally, using the maximum possible value of $AC(Q^*)$, one can determine the corresponding management-cost penalty.

## Extensions

The basic EOQ business scenario can be extended in a wide variety of ways. Here are just a few:

### *Integer and Case Quantities*

Equation (4) often results in a $Q^*$ that isn't an integer. For example, suppose that Lauren Worth computes $Q^* = 346.09$ units for some item she is analyzing. Since supplies can only be ordered in integer quantities, Lauren knows that Capital Hospital's supplier would laugh at her if she attempted to order 346.09 units. Given the shape of Fig. 8.1, it is obvious that Lauren's selected $\hat{Q}$ should be either the integer just above or just below the noninteger $Q^*$, since an even larger or even smaller integer value, respectively, would increase $AC(\hat{Q})$ unnecessarily. For example, given $Q^* = 346.09$, Lauren should order either $\hat{Q} = 346$ or 347 units. But which? Strictly speaking, Lauren *should* evaluate $AC(\hat{Q}= 346)$ and $AC(\hat{Q}= 347)$ and chose the $\hat{Q}$ with the smaller $\mathrm{AC}(\hat{Q})$.

However, given the fact that the cost-penalty ratio associated with a $\hat{Q}$ value different from $Q^*$ depends only on how far the ratio of $\hat{Q}/Q^*$ is from 1— see equation (7)— and that, whether we choose $\hat{Q}= 346$ or 347, the corresponding $\hat{Q}/Q^*$ ratio will be *very close* to 1, Lauren could simply round $Q^* = 346.09$ in the conventional manner (i.e., to $\hat{Q}= 346$) and pay a little or no cost penalty

Another practical consideration in purchasing scenarios is that suppliers may only accept orders for case quantities, where a case contains, say, 12 or 48 units.

Continuing the example above, suppose the EOQ formula prescribes $Q^* = 346.09$ units, but that this item is sold only in cases of 12. Hence, $Q^* = 346.09/12 = 28.84$ cases. Again, strictly speaking, Lauren should evaluate $AC(\hat{Q} = 28$ cases $= 336$ units) and $AC(\hat{Q} = 29$ cases $= 348$ units) and choose the $\hat{Q}$ with the smaller $AC(\hat{Q})$. Or, equivalently, Lauren should pick the $\hat{Q}$ whose $\hat{Q}/Q^*$ ratio is the closest to 1. However, since, again, both these ratios are very close to 1, Lauren *could* simply round conventionally—that is, round up to 29 cases—and pay a little or no cost penalty.

## Other Constraints on Q

In some scenarios there are other constraints on the order quantity. For example, in Cardinal Hospital's business scenario the supplier for some given item might require that some minimum quantity be ordered. To illustrate, suppose in the scenario in which $Q^* = 346.09$, the supplier requires a minimum order of 500 units. In such scenarios, Fig. 8.1 tells Lauren that the best thing to do is to order either $Q^*$ or the supplier's minimum amount, whichever is larger. Hence, in this example, Lauren should select $\hat{Q} = 500$ units (since ordering a larger amount will only yield an $AC(Q)$ larger than $AC(\hat{Q} = 500)$.

In some scenarios, storage considerations may constrain $Q$ to fit into the assigned storage space. In such scenarios, Fig. 8.1 tells us that the best thing to do is to order $Q^*$ or the capacity of the assigned space, whichever is smaller, since ordering an even smaller $\hat{Q}$ will yield higher costs.

## Quantity Discounts

In many purchasing scenarios suppliers offer discounts on the cost of the units themselves in order to motivate buyers to purchase in larger quantities. Such "quantity discount" scenarios are straightforward to handle. In particular, the goal remains the same as in the basic EOQ model: to select a value of $Q$, $Q^*$, that minimizes $AC(Q)$. However, since the unit cost, denoted c, in the basic EOQ scenario above, now depends on the order-quantity, $AC(Q)$ now must include the cost of a year's supply. In other words, equation (1) and not equation (2) should be used in evaluating $AC(Q)$.

As an illustration, consider the most common quantity discount, the so-called "all-units" discount scheme. Under this scheme if $\hat{Q}$ is below some "breakpoint quantity," denoted "b," then each unit of $Q$ costs a higher price, denoted $C_H$; however, if $\hat{Q}$ is greater than or equal to b, then each unit costs a lower per-unit price, denoted $C_L$.

The simplest way to handle this scenario is to determine the best choice of $Q$, first, with unit cost $C_H$, then, with unit cost $C_L$; and, finally, select the best overall $\hat{Q}$. Hence, Lauren would first compute $Q_H^*$ (that is, $Q^*$ if the unit cost is $C_H$), using equation (4). Then, based on equation (1), compute $AC(Q_H*)$. Next, compute $Q_L*$ (that is, the $Q*$ if the unit cost is $C_L$). However, in this step it is important to recognize that in order to qualify for the lower unit cost, $C_L$, $Q_L*$ must be greater than or equal to *b*, the breakpoint quantity. Hence, as described above, the order-quantity Lauren should select is the maximum of $Q_L*$ and b. Finally, Lauren should order either this quantity or $Q_H*$, depending on whose *AC*(*Q*)—using equation (1)—is lower.

## *Finite Planning Horizon*

Another major assumption of the EOQ business scenario is that demand goes on forever; in other words that the "planning horizon" is infinite. If, instead, management knows that demand will only last for, say, a fixed *T* years, then it would be foolish to order EOQ every time inventory fell to zero; and then, if inventory were leftover at the end of *T* years, to throw away those leftovers.

Schwarz (1972) has shown that, in this business scenario the optimal order-quantity is stationary (i.e., *Q** will never change); that is, management should order an equal quantity *n* times over *T* years. Schwarz further demonstrates that the optimal *n*, denoted *n** satisfies

$$n(n+1) \geq \frac{D \times h \times T^2}{2\mathrm{K}} \geq (n-1)n. \tag{8}$$

In order to develop some intuition about n*, note that, given (8), then $n*^2$ is approximately equal to

$$n^{*2} \approx \frac{D \times h \times T^2}{2\mathrm{K}} = \frac{D^2 \times T^2}{Q^{*2}},$$

which, in turn, means that n* is approximately equal to *DT*/*Q**; in other words, *n** is approximately equal to the total number of units demanded over the finite horizon, *DT*, divided by the number of units in an EOQ. Hence, a very good rule-of-thumb for picking n is to determine how many EOQs would satisfy *DT*, and then round this number to the closest integer.

## *(Q,r) Policies*

In order to gain insight into the order-quantity decision, the EOQ business scenario is structured so that the "order-point decision"—that is, what level of inventory, denoted r, should signal management to order its selected $\hat{Q}$?—is a trivial one. In its simplest form, the EOQ business scenario assumes that the order leadtime is zero. However, as long as demand (or usage) over the leadtime is known, the optimal order-point is to order whenever inventory equals a leadtime's supply.

Unfortunately, there are many real-world scenarios in which management *does not know* what demand during the leadtime will be! This leads to two important questions, one of them of some theoretical significance, and one of them of great practical significance.

The theoretical question is: How should the order-quantity and order-point denoted "(Q,r)," be chosen in order to minimize, average annual cost/time? The practical question is: How "good" are EOQ order-quantity decisions in business scenarios like this; that is, if one selected *Q* = *EOQ* and then selected a corresponding r, then would the cost-penalty be large or small?

The theoretical question has been addressed by several researchers, under a variety of different assumptions. For example, consider a scenario where all of the assumptions of the EOQ scenario apply, except that leadtimes are fixed and known, and uncertain customer demand is generated by a probability distribution such as the Poisson. For this scenario, Federgruen and Zheng (1992) describe an algorithm for computing (*Q**,*r**).

The practical question—which basically asks: How "good" is *EOQ* as the value of *Q* in a (*Q*,*r*) policy?— is an important, since, if the answer was that *EOQ* was "bad" at prescribing the order-quantity, then virtually every implemented real-world (*Q*,*r*) policy would be "bad," too, since EOQ is the de facto choice for the order-quantity in virtually every (*Q*,*r*) business setting.

The rationale for this de facto choice has been, as we have seen above, that inventory-management cost is insensitive to the choice of the order-quantity decision: insensitive to demand being known or constant, and even insensitive to its fixed ordering and inventory-holding costs being known or constant. Hence, as long as one picked an order-quantity in the "ballpark" of the correct value, one would pay a small cost penalty. The expectation, then, was that something like the same insensitivity applies in selecting the *Q* in a (*Q*,*r*) policy; i.e., that EOQ would provide an order-quantity in the ball park.

Zheng (1992) brilliantly validates the use of EOQ as a very good choice for *Q* in one specific (*Q*,*r*) scenario. Although the details are beyond the scope of this chapter, we will quote and interpret two of Zheng's findings: First, "… in contrast to the (basic) EOQ model, the controllable costs of the stochastic model due to the selection of the order quantity (assuming the (order point) is chosen optimally for every order quantity) are actually smaller…" What this means is that the cost/time in the (*Q*,*r*) model is *less* sensitive to the order-quantity decision than the cost/time in the EOQ model. Second, Zheng is able to prove that "…the relative increase of the costs incurred by using the (order-quantity) determined by the EOQ instead of the optimal from the stochastic model is no more than 1/8, and vanishes when ordering costs are significant relative to other costs." What this means is that the cost-penalty ratio associated with selecting *Q* to be *EOQ* and then selecting the order-point optimally is no more than 1.125; that, at worst, management would incur a 12.5% management-cost penalty as a consequence of its incorrect choice of *Q*.

Gallego (1998) describes heuristic (*Q*,*r*) policies and provides additional references.

### *Net-Present Value Criterion*

The EOQ business scenario uses the criterion of minimizing the average cost/time over an infinite time horizon in determining the optimal order quantity. One alternative to this criterion is net-present value. As applied in the EOQ scenario, we wish to select the order-quantity $\hat{Q}$ that minimizes the net-present value (i.e., discounted total cost) associated with this order quantity.

There are three cost inputs: (1) the fixed order cost *K*; and (2) the cost of capital, denoted i, on the dollars invested in the units, and measured in dollars/(dollar × time); and (3) the cost of the units themselves, denoted *c*, as above, and measured in dollars/unit.

It can be shown, see Zipkin (2000) for example, that the optimal Q in this scenario, denoted "$Q^*_N$," is approximately equal to EOQ; that is,

$$Q_N^* \approx \sqrt{\frac{2K \times D}{i \times c}}.$$

## Application: Cash Management

The EOQ model can be applied to *any* quantity decision that is repeated over time, where there is a trade-off between a fixed cost (i.e., EOQ'*s* order cost *K*) and a variable cost associated with that decision (i.e., EOQ's inventory-holding cost), and where the choice is based on minimizing the cost/time associated with that decision. The most popular such application involves the management of cash.

The simplest version of a cash-management scenario corresponds to the basic EOQ scenario, but with its inputs reinterpreted. In particular: (1) *D* is the rate at which cash is being accumulated from some business activity into a noninterest bearing account (e.g., checking account); (2) *h* is the interest being paid in an interest-bearing account (e.g., money-market account), and is measured in dollars/(dollar x time); and (3) *K* is the fixed cost incurred in transferring any amount of money, *Q*, from the non-interest to the interest-bearing account. Hence, *Q** from (4) is the optimal transfer quantity.

## Historical Background

The EOQ model and formula are attributed to Ford Whitman Harris (Harris, 1913), and originally published in 1913 in *Factory, The Magazine of Management*. However, despite this magazine's wide circulation among manufacturing managers, Harris' article had been lost until it was found by Erlenkotter (See Erlenkotter 1989).

In the years between Harris' article being lost and found, the EOQ model, and, in particular, the square-root equation, (4), was attributed to others, some calling it the "Wilson lot size formula," others calling it "Camp's formula." Erlenkotter (1989) describes the early development of the EOQ literature and sketches the life of Harris as an engineer, inventor, author, and patent attorney.

**Building Intuition** The EOQ formula provides near-optimal order quantities—that is, order quantities that have small management-cost penalties—even in realistic management scenarios when the unrealistic assumptions of the EOQ model don't apply. In particular, the EOQ formula will provide near-optimal order quantities regardless of whether the item's demand rate is either known or constant and regardless of whether or not its associated ordering and inventory-holding costs are either known or constant.

More broadly, the notion of a management-cost penalty introduced in this chapter is an important aid in thinking strategically about the adoption of *any* management system. That is, if we take it as granted that better information and more sophisticated decision-making systems yield better performing (i.e., lower cost) operations, then question is not: *When* or *how* to acquire/ adopt them? Instead, the question is: *Are these more sophisticated systems worth it?*

## References

Erlenkotter, D. (1989) "An Early Classic Misplaced: Ford W. Harris's Economic Order Quantity Model of 1915," *Management Science* 35:7, pp. 898–900.

Federgruen, A. and Y. S. Zheng. (1992) "An Efficient Algorithm for Computing an Optimal (r,Q) Policy in Continuous Review Stochastic Inventory Systems," *Operations Research* 40:4, pp. 808–813.

Gallego, G. (1998) "New Bounds and Heuristics for (Q,r) Policies," *Management Science* 44:2, 219–233.

Harris, F. M. (1913) "How Many Parts to Make at Once," *Factory*, *The Magazine of Management* 10:2, 135–136, 152. Reprinted in *Operations Research* 38:6 (1990), 947–950.

Lowe, T. J. and L. B. Schwarz. (1983) "Parameter Estimation for the EOQ Lot-Size Model: Minimax and Expected-Value Choices," *Naval Research Logistics Quarterly*, 30, 367–376.

Schwarz, L. B. (1972) "Economic Order Quantities for Products with Finite Demand Horizons," *AIIE Transactions* 4:3, 234–237.

Zheng, Y. S. (1992) "On Properties of Stochastic Inventory Systems," *Management Science* 38:1, 87–103.

Zipkin, P. (2000) *Foundations of Inventory Management*, McGraw-Hill Higher Education, New York, New York.

# Chapter 9
# Risk Pooling

**Matthew J. Sobel**
**Case Western Reserve University**

*This chapter shows that the following simple idea can be applied in many ways to manage business risks in the face of uncertainties. The standard deviation of a sum of interdependent random demands can be lower than the sum of the standard deviations of the component demands.*

## Introduction

At the end of an exciting and tension-filled week at Salmon Pools, Inc., Richard and Elizabeth congratulated each other on Salmon Pools' acquisition of Maple Leaf Pools (MLP). MLP, a Canadian manufacturer of accessories for above-ground swimming pools, was headquartered in Montreal. As Rich returned to his office, he considered the vexing problem of integrating the operations of Salmon Pools and MLP. Salmon Pools (SP), a medium-sized manufacturing company headquartered outside Springfield in central Massachusetts, was a major North American manufacturer and distributor of above-ground swimming pools and accessories. Liz was President and CEO, and Rich was Executive Vice President and oversaw purchasing, product design, manufacturing and distribution.

MLP was an attractive acquisition because its product lines complimented those of SP with little overlap. Although Liz and Rich had already decided to continue production activities both in Montreal and Springfield, the following issues were unresolved.

1. Both companies' products used the same kinds of raw materials and purchased parts such as specialty steel, small motors, plastic sheets, and pressure-treated wood. What were the economic advantages and disadvantages of maintaining separate inventories of the same items in Springfield and Montreal?
2. SP and MLP used different vendors for their purchases. Was it economical for the consolidated firm to *dual-source*, i.e., to purchase the same item from two (or more) suppliers?

Rich decided to ask Anne, SP's newly-hired OP (Operations Planner), to discuss these issues with him. Anne had just completed a Master's Degree in Operations Research

D. Chhajed and T.J. Lowe (eds.) *Building Intuition*: *Insights From Basic Operations Management Models and Principles*.
doi:10.1007/978-0-387-73699-0, © Springer Science + Business Media, LLC 2008

with an emphasis on operations management, and Rich hoped that her coursework included topics that would help him answer these questions. Also, he thought that it would be an effective way for her to learn more about SP's operations. After Rich briefly outlined the issues to Anne, they scheduled a meeting a few days later. Until then, she would prepare suggestions for analyzing the economic trade-offs of inventory centralization. Depending on the outcome of that meeting, he would ask her to do the same thing for dual sourcing. When she left his office she was elated. She realized that these decisions would be important to the future of the consolidated company and she knew that she could use the *square root law of inventory centralization* that she had learned in class.

## The Square Root Law of Inventory Centralization

At the meeting, Anne began with the basic single-location Economic Order Quantity model that is often called the *EOQ* model.[1] Suppose that an electric motor is used at the rate of 20 motors per week; so its annual consumption is $r$ = 1,000 motors per year (the factory operates 50 weeks per year). Its holding cost is $h$ = $10 per motor per year and its ordering cost is $K$ = $100 each time an order is placed. She explained that the holding cost includes opportunity cost, breakage in storage, theft, and property taxes. The ordering cost includes the expense of placing the order with the vendor, handling the delivery when it arrives, putting it in storage at SP (or MLP), and updating the inventory records to reflect the delivered items that were now in storage.

Rich said that he had learned about the EOQ model. Anne acknowledged that he already knew that the periodic order quantity, $Q$, that would minimize the sum of the annual holding and ordering costs was

$$Q^* = \sqrt{2rK/h} = \sqrt{2\times 1,000\times 100/10} = 141. \tag{1}$$

That is, if $Q$ motors were ordered every $Q$/20 weeks (or $Q/r$ years), then the annual holding and purchasing costs (not including payments to the vendor) would be

$$(r/Q)K + (Q/2)(Q/r)h. \tag{2}$$

This total cost is minimized by $Q^*$ given by (1) and the resulting sum of the annual holding and purchasing costs in (2) is $\sqrt{2Krh}$

Anne observed that if the costs and the weekly demand of this motor are the same in Montreal as they are in Springfield, and separate inventories are maintained in both locations, then by using the EOQ ($Q^*$) at both locations, the sum of the annual ordering and holding costs at the two locations will be

$$2\sqrt{2Krh} = 2\sqrt{2\times 100\times 1,000\times 10} = \$2,828 \text{ per year.} \tag{3}$$

[1] Read Chap. 8 to learn more about the Economic Order Quantity model.

Then she asked, instead of maintaining separate inventories at both factories, suppose that the needs at both places were met from a single consolidated inventory. Then the cost per order and the annual holding cost rate would remain the same as before, but the annual usage would rise to 2*r* motors per year. So the EOQ at the consolidated inventory location would rise from $Q^* = 141$ to

$$Q^{**} = \sqrt{2(2r)K / h} = \sqrt{2\times 2{,}000\times 100 / 10} = 200, \tag{4}$$

and the annual ordering and holding cost would drop from $2\sqrt{2Krh} = \$2{,}828$ to

$$\sqrt{2K(2r)h} = \sqrt{2\times 100\times 2{,}000\times 10} = \$2{,}000 \text{ per year.} \tag{5}$$

Comparing (3) and (5), the ratio of the minimal annual ordering and holding costs with consolidated inventories to the same costs with separate inventories is

$$1/\sqrt{2}. \tag{6}$$

Instead of consolidating two inventories, if *n* separate and identical inventories were consolidated at a single location, then the ratio would be $1/\sqrt{n}$ . This ratio explains the label the *square root law of inventory centralization*.

Anne commented that she had learned in class that the economics of centralization versus decentralization were more complex than these calculations suggested. For example, if a sizeable inventory were held only at one factory, then there would be shipping costs to transport the motors frequently from that factory to the other factory. On the other hand, the consolidated purchasing process might have more market power than either factory could have alone, and it might be able to negotiate a lower unit cost from a vendor. If trucks would frequently drive from one factory to the other, regardless of whether inventories were consolidated or not, then the incremental logistics cost might be negligible and the net economic improvement due to centralization might be greater than $1/\sqrt{2}$.

Rich noted that the factor of $1/\sqrt{2}$ applied to a quantity that included the square root of the ordering cost, *K* [see (1), (3), (4), and (5)], so the net advantage would not be important if *K* were small. He explained to Anne that years ago SP and MLP had each decided to specialize in high quality service niches of their industries. Each of them had structured its operations to respond to customers' needs very quickly. Although this obliged them to maintain high inventories of finished goods, they retained their largest customers year-after-year and they were often able to charge higher prices than the competition. It also meant that they had built relationships with suppliers of raw materials and parts who could respond very quickly to changing needs at SP and MLP. Rich noted that one of the side effects of *rapid response* was that both SP and MLP had partnered with suppliers to implement new procedures and technologies that lowered ordering costs. In other words, *K* in the EOQ was quite small for most items that were inventoried either in Springfield or Montreal.

Finally, Rich noted that the reduction in annual ordering and holding costs, that is the difference between (3) and (5), would be small if the ordering cost, *K*, were low.

Let $\Delta$ be the difference. Then

$\Delta^2 = \left[2\sqrt{2Krh} - \sqrt{2K(2r)h}\right]^2 = Krh4\left(3 - 2\sqrt{2}\right)$, and $\Delta$ would be small if the ordering cost, *K*, were low. In that case, would the economic advantage of consolidating inventories depend primarily on the trade-off between higher logistics costs and lower purchasing costs? He suggested that Anne consider these newly learned facts about the business, and that they resume their discussion of inventory consolidation in a few days. Anne was obviously downcast as she left the meeting room. She had pinned her hopes for making a big contribution on an inappropriate intellectual framework.

Later that day, Richard saw Elizabeth, Salmon Pool's CEO, and she asked him whether he had reached any conclusions about inventory consolidation and dual-sourcing. He said that Anne was inexperienced at SP, but she was learning quickly and he expected to be able to give Elizabeth more information after his next meeting with Anne.

## Safety Stock

At the next meeting, Anne began by explaining that she had learned that inventory was often held as a hedge against uncertainty. Since customer demand fluctuates from week-to-week, you cannot be sure exactly when to reorder an item so that the delivery arrives just before a stockout occurs. *This uncertainty leads to additional economic advantages of inventory consolidation.*

For example, suppose that successive weeks' demands for the electric motor at SP and MLP are independent random vectors

$(D_{S1}, D_{M1}),(D_{S2},D_{M2}),\ldots$ and that for each week *t*, $(D_{St}, D_{Mt})$ has the same probability distribution as $(D_S, D_M)$. Let $\sigma_S{}^2$ and $\sigma_M{}^2$ be the variances of a generic week's demands $D_S$ and $D_M$, respectively, and let $\rho$ be the correlation between $D_S$ and $D_M$. Anne thought that the weekly demands were unlikely to be independent because demand for pools and accessories is driven partly by the weather, and the weather is sometimes similar in Montreal and Springfield. In any case, the variance of the weekly demand at the consolidated inventory location would be $\sigma_{\mathrm{T}}^2$ which is given by

$$\sigma_{\mathrm{T}}{}^2 = \sigma_{\mathrm{S}}{}^2 + \sigma_{\mathrm{M}}{}^2 + 2\rho\sigma_{\mathrm{S}}\sigma_M. \tag{7}$$

For purposes of illustration, from now on assume that $D_S$ and $D_M$ have the same variance, say $\sigma^2{}_M = \sigma^2{}_S = \sigma^2$ so the variance of the consolidated weekly demand would be

$$\sigma_T{}^2 = 2\sigma^2(1+\rho). \tag{8}$$

This variance would be between zero and $4\sigma^2$ because $\rho$, the correlation coefficient, can be any number between −1 and +1.

**Building Intuition**

Centralizing inventories sometimes yields great economic benefits because the safety stock can be reduced. If the demands at the decentralized locations are negatively correlated, then the safety stock at a centralized location can be considerably less than the combined safety stocks at separate locations *while providing the same stockout protection!* The safety-stock advantage is nil if the demands are strongly positively correlated, but the advantage grows as the correlation shifts from strongly positive to strongly negative.

The importance of the variance is that it affects the *reorder point, R*. Anne explained that many inventory systems (like those at SP and MLP) have up-to-date information on each item's inventory levels, and they trigger the replenishment of an item when the *inventory position* drops to a trigger point, *R*, called the *reorder point*. An item's inventory position is the number of units in stock minus the number of units backordered (if any) plus the number of units contained in replenishment orders that are on the way from suppliers but have not yet been delivered. The reorder point should be high enough so that the demand during the procurement or production *lead time L*, is usually less than *R*. She understood that the rapid response programs at SP and MLP meant that the lead times were reliable, so she would treat the lead time *L* as a number rather than a random variable.

Two criteria are widely used to choose a reorder point. A *safety level criterion* selects *R* so that the risk of a stockout during the lead time, that is the probability that $\sum_{t=1}^{L} D_t > R,$ is a small fraction labeled $\alpha$; say $\alpha = 0.025$. The rationale for this criterion is that stockouts occur only during lead times, so the inventory manager should keep the risks low during lead times. A *service level criterion* focuses on customers who, after all, want good service and are not particularly interested in the minutia of lead times or reorder points. So this criterion selects the reorder point, *R,* to yield a high *fill rate*. The *fill rate*, often denoted $\beta$, is the fraction of customer demand that is filled immediately from on-hand inventory. The choice of $\beta$ should reflect the importance of not running out of stock, and typical values are between 0.90 and 0.99. Anne noted that these criteria had economic consequences due to inventory levels and customer patronage, and Rich would no doubt want to see the tradeoffs before he chose $\alpha$ or $\beta$.

## Safety Level Criterion for Safety Stock

If Rich opted for a safety level criterion and the inventories remained unconsolidated, then *R* at Springfield (SP) would be chosen so that

$$\alpha = P\{\sum_{t=1}^{L} D_{St} > R\}. \tag{9}$$

She said that the demand processes at many inventory systems were analyzed with a normal distribution. Let $Z$ be a standard normal random variable[2] and let $\Phi(\cdot)$ be its distribution function. Then $\dfrac{\sum_{t=1}^{L} D_{St} - L\mu}{\sigma_S \sqrt{L}}$ is a random variable with (approximately) the same distribution function as $Z$, namely $\Phi(\cdot)$. So (9) leads to

$$\alpha = P\{\sum_{t=1}^{L} D_{St} > R\} = P\{\frac{\sum_{t=1}^{L} D_{St} - L\mu}{\sigma\sqrt{L}} > \frac{R - L\mu}{\sigma\sqrt{L}}\}$$
$$= P\{Z > \frac{R - L\mu}{\sigma\sqrt{L}}\} = 1 - \Phi\left(\frac{R - L\mu}{\sigma\sqrt{L}}\right).$$

The expression

$$\alpha = 1 - \Phi\left(\frac{R - L\mu}{\sigma\sqrt{L}}\right) \tag{10}$$

can be evaluated easily with widely available tables of $\Phi(\cdot)$. Let $z$ be the place at which the standard normal random variable's cumulative area is 1 - $\alpha$ to the left and $\alpha$ to the right. That is, $\alpha = P\{Z>z\}=1-\Phi(z)$.

Then (10) corresponds to

$$z = \frac{R - L\mu}{\sigma\sqrt{L}} \text{ which is } R = L\mu + z\sigma\sqrt{L}. \tag{11}$$

For example, if $\alpha = 0.025$, $L$=1, and σ = 4 then z = 1.96 and $R$ = 1(20) + (1.96)(4)(1) = 27.84.

The *safety stock*, the hedge against uncertain demand during the lead time, is $z\sigma\sqrt{L}$, so in this example it is 7.84 units. If there are separate inventories, the cumulative safety stock level at SP and MLP is

$$2z\sigma\sqrt{L}. \tag{12}$$

If the inventories are consolidated, the lead time is still $L$ = 1, the mean consolidated demand per week is $\mu = 2\times 20 = 40$ and $\alpha$ is still 0.025 so $z$ = 1.96. However, the variance of demand during the lead time, from (8), is $\sigma_T{}^2 = 2\sigma^2(1+\rho) = 32(1+\rho)$.

[2] That is, z is a normal random variable with mean zero and variance one.

Applying (10) to consolidated demand,

$$R = L\mu + z\sigma_T\sqrt{L} = (1)(40) + (1.96)(\sqrt{32(1+\rho)}\sqrt{1} = 40 + 11.09\sqrt{1+\rho}.$$

So the safety stock, $11.09\sqrt{1+\rho}$, will range from 0 to 15.7 as the correlation between demands at the two locations varies from –1 to + 1. If the demands are uncorrelated ($\rho = 0$ ), the safety stock would be 11.09. Generally, the safety stock is

$$z\sigma_T\sqrt{L} = z\sigma\sqrt{L}\sqrt{2(1+\rho)}. \tag{13}$$

Contrasting (12) and (13), the ratio of the safety stocks in the centralized versus decentralized systems is

$$\sqrt{\frac{1+\rho}{2}}. \tag{14}$$

Anne observed that expressions (6) and (14) share the factor $1/\sqrt{2}$ which is frequently encountered when activities are aggregated. Expressions typically include a factor of $1/\sqrt{n}$ when *n* activities are consolidated. She then turned to the correlation coefficient, $\rho$ , in (14). Unlike (6) which is based on the presumption of deterministic demand, (14) quantifies the advantage of *pooling the risks* of random demands. The factor $\sqrt{1+\rho}$ in (14) often arises when risks are pooled. That is, the correlations among pooled risks affect the economic benefits. *The economic advantage of pooling is greatest if the risks (demands) are opposed (strongly negatively correlated,* $\rho = -1$) *, and the advantage is nil if the risks reinforce each other (strongly positively correlated,* $\rho$ = +1).

Anne commented that the differences in safety stocks could yield substantial savings to the consolidated operations of SP and MLP. If demands are uncorrelated, then the sum of the decentralized safety stocks is 7.84 × 2 versus $7.84\times\sqrt{2(1+\rho)}$. Multiplying the difference by *h* = $10, the annual unit holding cost, yields $78.4(2-\sqrt{2(1+\rho)}) = 110.86(\sqrt{2}-\sqrt{1+\rho})$.

If the demands are uncorrelated, this would save nearly $46 per year. SP and MLP have many hundreds of items in their inventories, and these savings could add up!

Anne started to explain that similar conclusions would be reached if they chose the reorder point to ensure a high fill rate. At this point, Richard’s concentration ebbed and his eyes started to glaze. He noted that SP and MLP were proud of their fill rates and a service level criterion was probably more appropriate than a safety level criterion. Even so, he said that many stocked items are resupplied periodically, say weekly, and the essential decision was the size of the shipment. That is, SP uses a *base-stock level* policy to manage inventories of many items rather than a reorder point-reorder quantity policy. He thought that MLP did the same. Rich commented that Anne had given him useful ideas to ponder and suggested that they meet again in a few days to review the benefits of consolidating inventories that are managed with base-stock level policies.

Later that afternoon, Elizabeth, the CEO, asked Rich about the progress of his considerations of whether to maintain inventories in Montreal and Springfield or to consolidate inventories at one location. He said that his meetings with Anne were productive and soon he expected to set a date on which they would know which inventoried items to consolidate.

## Base-Stock Level Policies

At the following meeting, Anne started to summarize the *Newsvendor Model* but Rich interrupted to say that he had learned all about it in his executive MBA program.[3] He had found it to be a useful metaphor for economically balancing the risk of too many versus the risk of too few. Anne replied that now they could apply newsvendor models to calculate base-stock inventory target levels. She would use the same electric motor as an example, but now she realized that SP and MLP contacted their suppliers each Thursday afternoon to specify the numbers of motors to deliver the following Monday morning. The quantity of each order was determined so that the inventory position late on Thursday plus the quantity delivered on Monday would reach a target level, τ, called the *base-stock level*. If *I* is the inventory position late Thursday, then the quantity ordered is $\tau - I$. How should τ be selected, and how would it be affected by inventory consolidation?

Anne reminded Rich of the equation for the value of $\tau$ that maximizes the expected net profit in the single-period newsvendor model. Let *D* be the weekly demand for motors, a random variable, and let *F* be its distribution function. Further, assume that any demand that exceeds supply is backordered (customers get "rainchecks"), with *p* as the unit cost of backordering, and *h* as the cost of holding of holding a unit of inventory for a week. Rich interjected "Back orders are a really big deal and we work harder than you can imagine to avoid them. They cost a lot of money in extra shipping cost (which can be the entire profit) and order processing (backorders are expensive to process due to the number of steps that the paperwork goes through and the higher incidence of lost or mistaken orders). Also, lost goodwill and reputation is a BIG deal on back orders in a tight seasonal business. We have to pay it a lot of attention. For example, all the other stuff that is available to ship is possibly junk to the dealer (customer) if they cannot go out and install the product without a missing motor—so we can't ship any of it at all!" Anne, now more accustomed to his vehemence, said "I guess that *p* would be quite large relative to *h*." Continuing, she pointed out that τ is the solution to

$$F(\tau) = \gamma \qquad \text{where} \qquad \gamma = \frac{p}{p+h}. \tag{15}$$

[3] Chapter 7 in this volume presents the basic principles of the newsvendor model.

So if $p$ is large relative to $h$, then $\gamma$ is close to one and $\tau$ must be a large number to satisfy (15). In particular, if $D$ is normally distributed with mean $\mu$ and variance $\sigma^2$, then (15) corresponds to

$$\tau = \mu + z\sigma, \tag{16}$$

where $\alpha = 1 - \gamma = h/(h + p)$. For example, if $\mu = 20, \sigma = 4$, $h = 0.2$, $p = 1.8$, then $\gamma = 0.9$ and $z = 1.28$. So

$$\tau_D = 20 + (1.28)(4) = 25.12, \tag{17}$$

where $\tau_D$ refers to the target inventory level in a decentralized system. Therefore, if the same parameters were appropriate at MLP, the aggregate target levels at the decentralized locations each Monday morning would be 2 × 25.12 = 50.24. This total is

$$2\tau_D = 2\mu + 2z\sigma. \tag{18}$$

How does this total compare to the appropriate target inventory level in a consolidated inventory system?

The mean weekly demand would be $2\mu$ and the same numerical values of p and h should be used. But the variance of weekly demand, from (8), would be $2\sigma^2 (1 + \rho)$. So the base-stock level of the consolidated inventory system, $\tau_C$ , should be

$$\tau_C = 2\mu + 2z\sigma\sqrt{2(1+\rho)}. \tag{19}$$

If the parameters are $\rho = 0$, $\mu = 20, \sigma = 4, h = 0.2$, and $p = 1.8$, then $\alpha = 0.1$ and $z = 1.28$, so the consolidated base-stock level is $40 + (1.28)(4)\ \sqrt{2}\ = 47.24$. The difference between (18) and (19) is

$$\sigma z\sqrt{2}(\sqrt{2} - \sqrt{1+\rho}). \tag{20}$$

So the consolidated target inventory level is at least as low as the sum of the unconsolidated levels, and the difference gets larger as the correlation between demands at the two locations becomes more negative. Once again, *the economic advantage of pooling is greatest if the risks (volatility of demands) are opposed (strongly negatively correlated,* $\rho = -1$), *and the advantage is nil if the risks reinforce each other (strongly positively correlated,*$\rho = +1$*)*. With the illustrative values of the parameters, the difference is 3.

## Base-Stock Level Policies with Lost Sales

Rich noted that these calculations all presumed that demand was backordered when it exceeded supply. However, both SP and MLP competed in a high service niche of their industry and they were much more likely to lose a sale than to be able to issue a “rain check” when demand exceeded supply. What were the economic effects of consolidation if excess demand was lost, i.e., what happened in the *lost sales case*? Furthermore, how could they systematically consider the lost goodwill and damaged reputation when a stockout occurs?

**Building Intuition**

The economic advantage of pooling risks is greatest if the risks are opposed, that is if they are strongly negatively correlated. The advantage is nil if the risks are strongly positively correlated, and the advantage grows as the correlation shifts from strongly positive to strongly negative.

Anne answered that the lost sales and backordering results were similar. In both cases, high values of *p* should reflect the effects of lost goodwill and damaged reputation if a stockout occurs. Instead of (15), lost sales results in the following equation for the base-stock level that minimizes the long-run average cost per week:

$$F(\tau) = \frac{p-c}{p+h-c}. \tag{21}$$

The new parameter, *c*, the unit cost of a motor, is the price that the motor manufacturer charges. This parameter appears in the expression because the unit cost of excess demand is offset by not having to pay the manufacturer for the motor. If demand is normally distributed, then (16) remains valid with a different calculation of $\alpha$:

$$\alpha = \frac{h}{p+h-c}. \tag{22}$$

For example, if the unit cost is $c = 1$ and we use the parameters on which (17) is based ($\mu = 20$, $\sigma = 4$, $h = 0.2$, and $p = 1.8$), then $\alpha = 0.2$ so z = 0.84 and

$$\tau_D = 20 + (0.84)(4) = 23.6. \tag{23}$$

Similarly, (19) and (20) are valid with the altered specification (22) instead of (15). So the difference between the aggregate unconsolidated target inventory levels and the consolidated target level would be

$$\sigma z\sqrt{2}\ (\sqrt{2}\ -\ \sqrt{1}\ +\rho) = (4)(0.84)\sqrt{2}\ (\sqrt{2}\ - 1) = 1.97. \tag{24}$$

## Base-Stock Level Policies with A Fill Rate Criterion

Rich commented that he knew how they should calculate *h*, the unit holding cost for an item that they inventoried, but it was very hard to estimate *p*, the unit cost of excess demand. He knew that they might lose all of a customer’s business due to a stockout. How should he translate that risk to a numerical value of $p$? He thought that a fill rate was more appropriate for SP and MLP, and they achieved at least 95% on most items such as the motor, i.e.,$\beta = 0.95$. What were the advantages of consolidating inventories that were managed with base-stock level policies that responded to a fill rate criterion?

Anne said that the resulting equation for the base-stock level was more complicated than (15) but there was an excellent approximation if the succession of weekly demands were independent normal random variables with the same mean:

$$\tau \cong (L+1)\mu + \sigma\sqrt{L+1}\;\Phi^{-1}[\beta + \Phi(-\sqrt{L}\mu/\sigma)]. \tag{25}$$

In (25), $\Phi^{-1}$ is the function inverse of the standard normal distribution function $\Phi$. That is, $\Phi^{-1}(x) = z$ corresponds to $\Phi(z) = x$. For example, using a table of $\Phi$, $\Phi^{-1}(0.5) = 0$ and $\Phi^{-1}(0.99) = 2.33$. In order to illustrate (25), suppose that the parameters for a motor that is stocked at a decentralized location are $\mu = 20$, $\sigma = 4$, $L = 1$, and $\beta = 0.95$. Then $\tau_D$, the target base-stock level that yields a fill rate of 95%, is

$$\begin{aligned} \tau_D &\cong (1+1)(20) + (4)\sqrt{1+1}\;\Phi^{-1}[0.95 + \Phi(-\sqrt{1}(20)/4)] \\ &= 40 + 4\sqrt{2}\Phi^{-1}[0.95 + \Phi(-5)] = 40 + 4\sqrt{2}\Phi^{-1}[0.95] \\ &= 40 + (5.656)(1.645) = 49.30. \end{aligned} \tag{26}$$

If the motor is stocked at both locations, the aggregate target base-stock inventory is 2 × 49.30 = 98.60. Instead, if the motors are stocked at a consolidated location, then $\tau_D$, the target base-stock inventory level, is given by (25) with $\mu$ replaced by $2\mu$ and $\sigma$ replaced by $\sqrt{2\sigma^2(1+\rho)}$ :

$$\tau_C \quad (L+1)(2\mu) + \sigma\sqrt{L+1}\sqrt{2(1+\rho)}\Phi^{-1}[\beta + \Phi(\frac{-2\mu\sqrt{L}}{\sigma\sqrt{2(1+\rho)}})]. \tag{27}$$

Using the same parameter values as in (26), if demands at the two locations are uncorrelated, then

$$\tau_C = (2)(1+1)(20) + \sqrt{2}(4)\sqrt{2}\Phi^{-1}[0.95 + \Phi(\frac{-\sqrt{2}(20)\sqrt{1}}{4})] = 93.16. \tag{28}$$

Since the lead time is one week ($L = 1$), in the comparison of $\tau_D$ with $\tau_c$, the mean demand during the current week and the following week when the ordered goods arrive is 80 units. The remainder is a cushion for risk. That cushion is 98.60−80 = 18.60 if the inventories remain decentralized. The cushion is 93.16−80 = 13.16 if the inventories are consolidated and demands at the two locations are uncorrelated. Notice that $18.60/\sqrt{2} = 13.16$. Once again, consolidating two inventories yields improvements at a rate of $1/\sqrt{2}$ .

At this point Rich asked if consolidation would provide greater benefits if demands were negatively correlated. Anne confirmed his understanding by calculating the base-stock level when $\rho = -0.5$. The result was that $\tau_c = 89.30$ whose cushion of 9.30 would be a 29% reduction of the cushion with $\rho = 0$.

Rich closed the meeting by thanking Anne and saying that he understood the economic tradeoffs of inventory consolidation much more clearly than when they discussed it the first time. He said that the next steps were:

- To organize the information on hundreds of items that were stocked both at Montreal and Springfield in order to analyze the economic tradeoffs and finally reach decisions on which items should be stocked in a consolidated inventory system;
- To forecast the date by which the necessary information would be available so that consolidation decisions could be made;
- To understand when it was economical for the consolidated firm to *dual-source*, i.e., to purchase the same item from two (or more) suppliers.

He scheduled the next meeting and praised Anne! Later that day, he told Elizabeth, the CEO, that he could shortly tell her when the inventory consolidation decisions would be made.

## Dual Sourcing

When they met, Anne described the data that she would retrieve to estimate the potential savings from consolidating inventories. She intended to analyze the impacts of consolidation on inventory-related costs, payments to suppliers, and logistics costs. After she estimated the time she needed to obtain and organize the data, the discussion turned to dual sourcing.

Rich commented that industrialists all over the world were paying greater attention to multiple sources of supply since 2000. A fire in an Albuquerque semiconductor plant interrupted an important source of computer chips for two of the world’s major cell phone manufacturers, Nokia Corp. of Finland and Telefon AB LM Ericsson of neighboring Sweden. Global mobile phone sales were soaring and it was critical for both companies to find alternative suppliers promptly. While Nokia succeeded, Ericsson did not and lost at least $400 million in potential revenue.

**Building Intuition**

The economic advantage of dual sourcing can stem from pooling the risks of a supply disruption. The advantage is greatest if the risks are opposed, that is if they are strongly negatively correlated. The advantage is nil if the risks are strongly positively correlated, and the advantage grows as the correlation shifts from strongly positive to strongly negative.

Anne observed that an industry’s characteristics determined whether there were potential savings from awarding business to multiple suppliers. SP and MLP bought items that were made with standard production technologies so there were plenty of potential suppliers for most of them. Moreover, suppliers didn't depend on an SP order (or an MLP order) to move out on the "learning curve."

SP and MLP usually arranged for suppliers to ship items weekly or monthly during the production season. Purchasing an item from more than one supplier is *multiple sourcing*. Since SP and MLP were medium-sized companies, when they multi-sourced it was usually with two suppliers and they called it *dual sourcing*.

Anne said that the potential advantages to SP and MLP from dual sourcing depended on correlations of risks. They had already seen the importance of correlations in inventory consolidation. She said that she would illustrate these considerations with a small electric motor.

Each August, before the production season began, SP placed open orders for the motor with two suppliers. The production year's demand was estimated in October when half the estimated demand was ordered from each supplier; the finished motors were shipped two months later. There was a good reason why the motors were not ordered (and shipped) monthly between October and June. Although SP was a whale among manufacturers of above-ground pools, it was a minnow in the overall market for small electric motors. When the motor manufacturers received SP orders, they set up equipment that was unique to SP's needs and made the batch of ordered motors in a few weeks' time. The motor manufacturers used suppliers to make the corrosion-resistant casing that housed the motor. After a supplier completed the ordered batch of casings, the equipment that was specific to the SP casings was replaced with equipment for whatever job they did next. If the motors had been ordered monthly, then the cost of this setup and teardown would be absorbed by relatively few motors (one month's demand) and the unit cost of motors would have been too high.

The prices charged by the two motor suppliers were slightly different but their motors had the same quality. Did the risk of a supply disruption justify paying a higher price for half the motors? Label the suppliers $x$ and $y$ and let $c$ and $c + \delta$ be the prices charged by $x$ and $y$, respectively. Here, $\delta > 0$ indicates that $y$ charges a higher price than $x$. Let $X$ indicate whether supplier $x$ can supply the motors or has a supply disruption. That is, $X = 1$ if $x$ can supply the motors, and $X = 0$ if $x$ cannot supply the motors. Similarly, let $Y = 1$ if $y$ can supply the motors, and $Y = 0$ if $y$ cannot supply the motors. It is realistic to count on either supplier being able to provide *all* the motors if the other cannot provide any. Also, if both suppliers are unable to provide motors at the last minute, then SP can find another supplier who can ship fairly quickly but at a high price, say $K$ where $K > c + \delta$.

Let $\pi_x$ be the cost if SP sole sources with supplier $x$, and let $\pi_{xy}$ be the cost if SP dual sources. Let $\Delta$ be the expected difference between the cost of sole sourcing (with $x$) and the cost of dual sourcing; that is,

$$\Delta = E(\pi_X) - E(\pi_{XY}). \tag{29}$$

So dual sourcing is cost-effective if $\Delta \geq 0$. Under what conditions is $\Delta \geq 0$? The answer depends on the cost parameters $K$, $c$, $\delta$, and on the joint probability distribution of $X$ and $Y$.

The cost of sole sourcing $Q$ units is $cQ$ if $X = 1$, and it is $KQ$ if $X = 0$. So

$$\pi_X = cQX + KQ(1 - X). \tag{30}$$

The expression for $\pi_{\mathrm{xy}}$ is more complex; that cost is $(c + \delta/2)Q$ if $X = Y = 1$[4], it is $cQ$ if $X = 1$ and $Y = 0$, it is $(c + \delta)Q$ if $X = 0$ and $Y = 1$, and it is $KQ$ if $X = Y = 0$. Therefore,

$$\pi_{XY} = (\mathrm{c} + \delta/2)QXY + cQX(1 - Y) + (\mathrm{c} + \delta)QY(1 - X) + KQ(1 - X)(1 - Y). \tag{31}$$

Let $1-a$ be the probability of a supply disruption; that is, $a = P\{X = 1\} = P\{Y = 1\}$ and $1 - a = P\{X = 0\} = P\{Y = 0\}$. Using $a = E(X)$ in (30), the expected cost of sole sourcing is

$$E(\pi_x) = Q[K - a(K - c)]. \tag{32}$$

In order to specify $\Delta$, the difference between the expected costs with sole sourcing and multi-sourcing, we can use (31) and (32) after obtaining $E(\pi_{\mathrm{xy}})$. That task is complicated by terms in (31) involving $XY$. However, $E(XY)$ can be specified in terms of $a$ and the correlation coefficient of $X$ and $Y$. The correlation between $X$ and $Y$, labeled $\rho$, is

$$\rho = \frac{E(XY) - a^2}{a(1 - a)}$$

so $E(XY) = a^2 + \rho a(1 - a)$. Using this expression with (31) to specify $E(\pi_{\mathrm{xy}})$ leads to

$$E(\pi_{XY}) = Q\{K - 2a(K - c) + \delta a + (K - c - \delta/2)a[\rho(1 - a) + a]\}. \tag{33}$$

Subtracting (33) from (32) gives $\Delta$, the cost advantage of dual sourcing instead of sole sourcing:

$$\Delta = \{(K - c - \delta) - (K - c - \delta/2)[\rho(1 - a) + a]\}\, Qa. \tag{34}$$

This cost advantage is favorable ($\Delta \geq 0$) if

$$a + \rho(1 - a) \leq \frac{K - c - \delta}{K - c - \delta/2}. \tag{35}$$

Notice in (35) that the annual volume of motors, $Q$, plays no role in whether the cost advantage is favorable or not. Also notice in (33), (34), and (35) that dual sourcing is more favorable if the risks of supply disruptions from $x$ and $y$ are more strongly opposed. That is, the advantage grows as $\rho$, the correlation of $X$ and $Y$,

[4] The payments to $x$ and $y$, respectively, are $cQ/2$ and $(c + \delta)Q/2$, so the sum is $(c + \delta/2)Q$.

moves from +1 down to −1. Inventory consolidation had the same risk pooling influence of the correlation. Richard had become less attentive as the algebra seemed to dominate the presentation, so Anne gave him an example. Suppose $c = 10$, $K = 50$, $\delta = 0.5$, $a = 0.95$, and $\rho = 0.5$. With these parameters, the price of supplier *x* is 5% higher than the price of supplier *y*. The ratio on the left side of (35) is 0.975 and the right side is 0.995. This means that $\Delta \geq 0$; so sole sourcing is more expensive than dual sourcing.

Richard commented that $c = 10$ compared to $K = 50$ was unrealistic; he doubted that the price would rise five-fold even for an emergency shipment with a very short lead time. How low could *K* drop until it was no longer beneficial to dual source? Anne said that it was easy to rearrange (35) to specify indifference points. For instance, $\Delta \geq 0$ if

$$K \geq c + \left[1 + \frac{1}{(1-a)(1-\rho)}\right]\frac{\delta}{2}. \tag{36}$$

In the numerical example, the right side of (36) is 20.25, so it would be cost-effective to dual source if *K* were at least 20.25.

Similarly, (35) yields indifference points for the correlation coefficient, ρ, and the difference δ between the prices charged by *x* and *y*:

$$\rho \leq 1 - \frac{\delta}{2(1-a)(K-c-\delta/2)}, \quad \text{and} \quad \delta \leq \frac{2(1-a)(1-\rho)(K-c)}{1+(1-a)(1-\rho)}. \tag{37}$$

Using the parameters in the example, the right side of the inequality for ρ is 0.87, so dual sourcing would be cost-effective even if the correlation coefficient were significantly higher than 0.5. The right side of the inequality for δ is 1.95. So dual sourcing would be cost-effective until the higher priced supplier (*y*) charged 19.5% more than the lower priced supplier.

Richard thanked Anne for identifying some of the key considerations in a decision of whether or not to use multiple suppliers for the same goods or service. He said that her next project, after she had retrieved and organized the data on inventory consolidation, would be to organize data on dual sourcing at SP and MLP.

The next morning, when Rich described their progress to Elizabeth (the CEO), she asked whether risk pooling might assist them to obtain lower health insurance premiums for Salmon Pools employees. Most of the employees worked in Springfield, Massachusetts, but the firm also had employees in three US sales offices. Although employee benefits were the same at all geographic location, SP had separate insurance policies for the employees at each location.

Richard asked her why she didn't simply have the head of Human Resources (HR) ask the insurance broker (through whom the firm bought health insurance) if consolidating the policies would result in lower premiums. Elizabeth explained that she first wanted Richard's advice before intimating that the head of HR and the broker might not have protected the firm's interests adequately. Rich said that he would discuss with Anne whether the principles of risk pooling were as applicable to insurance premiums as to safety stocks.

## Pooling Insured Risks

Soon Anne and Richard met to discuss whether SP might be able to bargain for lower health insurance premiums if all employees were covered by one policy instead of four. She said that the underlying considerations for inventory consolidation were relevant, but the details were different. The square root law of inventory consolidation could not be used here, but the premise of an economy of scale remained valid. For example, an insurance company spends less money on billing SP and posting payments of premiums if there is one insurance policy instead of four. Rich replied that the insurance company's processing costs were not significantly different if there were one group policy or four. Anne agreed that differences in processing costs were probably unimportant, but there was an important economy of scale from the perspective of the insurance company.

Rich said that he did not see how consolidating the policies could affect the insurance company's total payments of health insurance claims. Anne agreed that the insurance company's total payout would be the sum of claims submitted by SP employees; that sum didn't depend on the manner in which employees were batched into policies. She observed, however, that risk pooling might provide the insurance company an important advantage that SP could exploit to obtain lower premiums.

> **Building Intuition**
>
> There are advantages from operating a business so that its *suppliers* can pool *their* risks to reduce *their* costs. Their cost reduction is greatest if their risks are opposed, that is if they are strongly negatively correlated. The cost reduction is nil if their risks are strongly positively correlated, and the cost reduction grows as the correlation shifts from strongly positive to strongly negative.

She illustrated the potential advantage with the effects of pooling the sales office employees at Mobile, Alabama and Salt Lake City, Utah. In order to use the same formulas as for inventory pooling, she let $D_M$ and $D_S$ denote next year's claims that would be paid to employees in Mobile and Salt Lake City, respectively.[5] If the risks are pooled into a single policy, then $\sigma^2$, the variance of the total payout under that policy, is given by (38):

$$\sigma_{\mathrm{T}}^{2} = \sigma_{\mathrm{S}}^{2} + \sigma_{\mathrm{M}}^{2} + 2\rho\sigma_{\mathrm{S}}\sigma_{M}. \tag{38}$$

In this formula, $\sigma_S$, $\sigma_M$, and $\rho$ are respectively the variance of $D_S$, the variance of $D_M$, and the correlation between them. In particular, for the remainder of this

[5] In Eq. (38) and the discussion of inventory consolidation, these symbols denote a week's demands at Springfield and Montreal, respectively.

discussion, suppose that $D_M$ and $D_S$ have the same variance, say $\sigma^2{}_{\mathrm{M}} = \sigma^2{}_{\mathrm{S}} = \sigma^2$; so (38) becomes (39):

$$\sigma_T^2 = 2\sigma^2(1+\rho). \tag{39}$$

The importance of the total variance, $\sigma_{\mathrm{T}}^2$, is that an insurance company is regulated by the insurance departments in the states in which it does business, and the company is required to hold some of its assets in a form that makes it very likely that the company can pay all of its claims next year. The requisite forms are low risk securities that are easily marketable (for example, US Treasury bills). In the discussion of a safety level criterion earlier in this chapter, $\alpha$ in (9) was the risk of a stockout during a replenishment leadtime. Here, $\alpha$ is the risk that the total health insurance claims exceed the amount of assets that the insurance company can convert to cash to pay those claims. Then the *safety reserve* is the amount of funds that the company must hold in the form of low-risk highly marketable securities, over and above the expected amount of the claims, in order to have the probability as high as 1 - $\alpha$ that it will be able to pay all of the claims.

There is a substantial opportunity cost attached to every dollar that the insurance company holds in safety reserve; that dollar cannot be placed in higher yield investments that are riskier or less liquid. So if $\sigma_{\mathrm{T}}^2$ is higher, the safety reserve is higher, the opportunity cost is higher, and the insurance company compensates by charging a higher premium.

Repeating (12) and (14)[6], if there are separate insurance policies for the employees at Mobile and Salt Lake City, the cumulative safety reserve level is[7]

$$2z\sigma. \tag{40}$$

If the employees at both locations are pooled into one policy, the safety reserve is

$$z\sigma_T = z\sigma\sqrt{2(1+\rho)}. \tag{41}$$

Contrasting (40) and (41), the ratio of the safety reserves in the pooled versus separate policies is

$$\sqrt{\frac{1+\rho}{2}}. \tag{42}$$

This ratio ranges from + 1, which occurs if the claims at the two locations are perfectly positively correlated ($\rho$ = +1), to 0, which occurs if the claims at the two locations are perfectly negatively correlated ($\rho$ = –1).

[6] Here, L = 1 in (12) and 13).

[7] Recall the notation z for the fractile of the standard normal distribution where 100α% of the area lies to the right.

Rich said that he doubted that the claims at the two locations were interdependent, so he wouldn't be surprised if $\rho$ were ~0. That would cause (42), the ratio of the safety reserves, to be $1/\sqrt{2}$ which is ~0.71. That is, pooling the employees would reduce the insurance company's safety reserve by about 29%. He thanked Anne for helping him understand why it might be profitable for the insurance company if Salmon Pools combined all its employees into a single health insurance policy.

Later that day, he made an appointment with Elizabeth to explain his insight. He suggested that their firm's director of HR contact the insurance broker and insurance company. Before renewing the insurance policies, they should analyze claims data to find out if consolidating the policies would increase the insurance company's profits (by reducing its opportunity costs). In that case, SP should negotiate to share the higher profits by paying lower health insurance premiums.

## Applications of Risk Pooling

Risk pooling is not a panacea and consolidation generates costs as well as benefits. The usual question is whether the benefits of consolidation sufficiently outweigh the added costs. For example, if inventories are held at numerous widely scattered depots that are close to customers, there are typically low costs to deliver the goods from the depots to the customers. On the other hand, having numerous depots may require a high cost to transport goods from factories to depots. Consolidating numerous depots into a single (or a few) central locations may reduce safety stocks with their associated opportunity costs. Although it will also reduce the costs of transporting goods from factories to depots, it will increase the costs of distributing goods from depots to customers. In some cases there will be a large net cost reduction but its magnitude may depend on the shrewdness of the selection of the centralized location. In other cases, net costs will increase. So it is essential to account for *all* cost changes before making a commitment to consolidate risks. Kulkarni et al. (2005) study the tradeoff between risk pooling and logistics costs in a multi-plant network.

The emerging field of financial engineering includes sophisticated methods to pool risks.

Much of the fable in this chapter concerns the aggregation of multiple locations; there is an opportunity to combine inventories at decentralized locations into a centralized depot at one location. *Delayed differentiation* is an application of risk pooling in which aggregation occurs *in time*. This arises, for example, when components are stocked after they are fabricated, but finished goods are assembled only after customer orders are received. There is a large and growing literature on this topic which is sometimes labeled *mass customization*.

Another operations management application of risk pooling is a firm's reduction of the number of different products it provides. A producing firm can reduce the variety of finished goods and services that it produces; a retail store can reduce the variety of goods that it stocks. In each instance, one should balance the lower costs due to a reduction in variety versus the lost revenue.

Production capability is yet another direction of aggregation. Here, a firm may be able to replace several specialized "machines" with a single "flexible" machine. Under what circumstances is flexible automation superior to a set of specialized capabilities? Some answers are provided by Graves and Tomlin (2003) and Tomlin and Wang (2005).[8]

Most research on risk pooling assumes that end-item demand fluctuates more slowly than production. However, some applications of risk pooling occur when the production of goods occurs in supply chains where there are significant delays between the date on which an additional output is sought, and the date on which it is finally available. These contexts are the *production-inventory systems* discussed by Benjaafar et al. (2005).

Dual sourcing is an important application of risk pooling. See the research reviews of dual sourcing by Elmaghraby (2000) and Minner (2003). Dual sourcing in a risk pooling context is discussed by Babich et al. (2007) and Tomlin and Wang (2005). See Latour (2001) for the consequences of the 2001 fire in a Albuquerque, N.M. semiconductor plant.

## Historical Background

Risk pooling is an old concept that came to operations management relatively recently. Five thousand years ago, Chinese sea-going merchants distributed their goods in several ships to reduce the risk of total loss. Insurance is a manifestation of risk pooling, and Lloyd's of London, the world's preeminent specialist insurance market, began in the seventeenth century.

Gary Eppen (1979) initiated the study of risk pooling in operations management. He considered the consolidation of inventories that are held in separate locations, and he showed that the inventory-related costs of the consolidated system are lower if the correlation between the correlated demands is more negative. His assumptions include normally distributed demand and the same parameters at each of the separate locations. However, Eppen's conclusions remain valid if the parameters are *not* the same at each location (Ben-Zvi and Gerchak 2005) or if demand has a nonnormal distribution (Chen and Lin 1989). Good brief surveys of the ensuing research are included in Benjaafar et al. (2005) and Gerchak and He (2003). The reader is directed to those papers for references to the material that was published between 1980 and 2003.

Aggregation in a supply chain raises the possibility of consolidating inventories that are maintained by different firms. The aforementioned literature considers the aggregate costs and benefits (*social welfare* in economics parlance) rather than the distribution of the costs and benefits among the participating firms. Two exceptions which take game theoretic approaches are Ben-Zvi and Gerchak (2005*)* and Hartman and Dror (2005*)*.

[8] Flexibility principles are presented in this volume in Chap. 3.

The material on base-stock level policies with a fill rate criterion is based on Sobel (2004).

**Acknowledgment** The author is grateful to Anne, Elizabeth, and Richard for all that they have taught him.

## References

Alfaro, J. A. C. and C. J. Corbett (2003) “The Value of SKU Rationalization in Practice (The Pooling Effect of Suboptimal Inventory Policies and Nonnormal Demand),” *Production and Operations Management* 12, 12–29.

Babich, V., P. H. Ritchken, H. and A. N. Burnetas (2007) “Competition and Diversification Effects in Supply Chains with Supplier Default Risk,” *Manufacturing and Service Operations Management*. 9, 123–146.

Benjaafar, S., W. L. Cooper and J. S. Kim (2005) “On the Benefits of Inventory Pooling in Production-Inventory Systems,” *Management Science* 51, 548–565.

Ben-Zvi, N. and Y. Gerchak (2005) “Inventory Centralization When Shortage Costs Differ: Priorities and Costs Allocation,” Technical Report, Dept. of Industrial Engineering, Tel-Aviv University.

Bertsimas, D. and I. Ch. Paschalidis (2001) “Probabilistic Service Level Guarantees in Make-to-Stock Manufacturing Systems,” *Operations Research* 49, 119–133.

Chen, M. S. and C. T. Lin (1989) “Effects of Centralization on Expected Costs in a Multi-Location Newsboy Problem,” *Journal of Operational Research Society* 40, 597–602.

Elmaghraby, W. (2000) “Supply Contract Competition and Sourcing Policies,” *Manufacturing & Service Operations Management*, 2, 350–371.

Eppen, G. D. (1979) “Effects of Centralization on Expected Costs in a Multi-Location Newsboy Problem,” *Management Science* 25, 498–501.

Gerchak, Y. and Q.-M. He (2003) “On the Relation Between the Benefits of Risk Pooling and the Variability of Demand,” *IIE Transactions* 35, 1027–1031.

Graves, S. C. and B. T. Tomlin (2003) “Process Flexibility in Supply Chains,” *Management Science* 49 (7), 907–919.

Hartman, B. C. and M. Dror (2005) “Allocation of Gains from Inventory Centralization in Newsvendor Environments,” *IIE Transactions* 37, 93–107.

Kulkarni, S. S., M. J. Magazine, and A. S. Raturi (2005) “On the Tradeoffs Between Risk Pooling and Logistics Costs in a Multi-Plant Network with Commonality,” *IIE Transactions* 37, 247–265.

Latour, A. (2001) “Trial by Fire: A Blaze in Albuquerque Sets Off Major Crisis for Cell-Phone Giants,” *The Wall Street Journal*, Jan. 29, A1.

Minner, S. (2003) “Multiple-Supplier Inventory Models in Supply Chain Management, A Review,” *International Journal of Production Economics* 81–82, 265–274.

Sobel, M. J. (2004) “Fill Rates of Single-Stage and Multistage Supply Systems,” *Manufacturing & Service Operations Management* 6, 41–52.

Tomlin, B. and Y. Wang (2005) “On the Value of Mix Flexibility and Dual Sourcing in Unreliable Newsvendor Networks,” *Manufacturing & Service Operations Management* 7, 37–57.

# Index

*Early Titles in the*

## INTERNATIONAL SERIES IN OPERATIONS RESEARCH & MANAGEMENT SCIENCE

(*Continued*)

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*
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/ *QUANTITATIVE MODELS FOR PERFORMANCE EVALUATION AND BENCHMARKING*
Ehrgott & Gandibleux/ *MULTIPLE CRITERIA OPTIMIZATION*: *State of the Art Annotated Bibliographical Surveys*
Bienstock/ *Potential Function Methods for Approx. Solving Linear Programming Problems*
Matsatsinis & Siskos/ *INTELLIGENT SUPPORT SYSTEMS FOR MARKETING DECISIONS*
Alpern & Gal/ *THE THEORY OF SEARCH GAMES AND RENDEZVOUS*
Hall/*HANDBOOK OF TRANSPORTATION SCIENCE - 2nd Ed.*
Glover & Kochenberger/ *HANDBOOK OF METAHEURISTICS*
Graves & Ringuest/ *MODELS AND METHODS FOR PROJECT SELECTION*: *Concepts from Management Science, Finance and Information Technology*
Hassin & Haviv/ *TO QUEUE OR NOT TO QUEUE*: *Equilibrium Behavior in Queueing Systems*
Gershwin et al/ *ANALYSIS & MODELING OF MANUFACTURING SYSTEMS*

## *Early Titles in the*
## INTERNATIONAL SERIES IN
## OPERATIONS RESEARCH & MANAGEMENT SCIENCE

(*Continued*)

Maros/ *COMPUTATIONAL TECHNIQUES OF THE SIMPLEX METHOD*
Harrison, Lee & Neale/ *THE PRACTICE OF SUPPLY CHAIN MANAGEMENT*: *Where Theory and Application Converge*
Shanthikumar, Yao & Zijm/ *STOCHASTIC MODELING AND OPTIMIZATION OF MANUFACTURING SYSTEMS AND SUPPLY CHAINS*
Nabrzyski, Schopf & Węglarz/ *GRID RESOURCE MANAGEMENT*: *State of the Art and Future Trends*
Thissen & Herder/ *CRITICAL INFRASTRUCTURES*: *State of the Art in Research and Application*
Carlsson, Fedrizzi, & Fullér/ *FUZZY LOGIC IN MANAGEMENT*
Soyer, Mazzuchi & Singpurwalla/ *MATHEMATICAL RELIABILITY*: *An Expository Perspective*
Chakravarty & Eliashberg/ *MANAGING BUSINESS INTERFACES*: *Marketing, Engineering, and Manufacturing Perspectives*
Talluri & van Ryzin/ *THE THEORY AND PRACTICE OF REVENUE MANAGEMENT*
Kavadias & Loch/*PROJECT SELECTION UNDER UNCERTAINTY*: *Dynamically Allocating Resources to Maximize Value*
Brandeau, Sainfort & Pierskalla/ *OPERATIONS RESEARCH AND HEALTH CARE*: *A Handbook of Methods and Applications*
Cooper, Seiford & Zhu/ *HANDBOOK OF DATA ENVELOPMENT ANALYSIS*: *Models and Methods*
Luenberger/ *LINEAR AND NONLINEAR PROGRAMMING, 2nd Ed.*
Sherbrooke/ *OPTIMAL INVENTORY MODELING OF SYSTEMS*: *Multi-Echelon Techniques, Second Edition*
Chu, Leung, Hui & Cheung/ *4th PARTY CYBER LOGISTICS FOR AIR CARGO*
Simchi-Levi, Wu & Shen/ *HANDBOOK OF QUANTITATIVE SUPPLY CHAIN ANALYSIS*: *Modeling in the E-Business Era*
Gass & Assad/ *AN ANNOTATED TIMELINE OF OPERATIONS RESEARCH*: *An Informal History*
Greenberg/ *TUTORIALS ON EMERGING METHODOLOGIES AND APPLICATIONS IN OPERATIONS RESEARCH*
Weber/ *UNCERTAINTY IN THE ELECTRIC POWER INDUSTRY*: *Methods and Models for Decision Support*
Figueira, Greco & Ehrgott/ *MULTIPLE CRITERIA DECISION ANALYSIS*: *State of the Art Surveys*
Reveliotis/ *REAL-TIME MANAGEMENT OF RESOURCE ALLOCATIONS SYSTEMS*: *A Discrete Event Systems Approach*
Kall & Mayer/ *STOCHASTIC LINEAR PROGRAMMING*: *Models, Theory, and Computation*
Sethi, Yan & Zhang/ *INVENTORY AND SUPPLY CHAIN MANAGEMENT WITH FORECAST UPDATES*

*Early Titles in the*

**INTERNATIONAL SERIES IN**

**OPERATIONS RESEARCH & MANAGEMENT SCIENCE**

(*Continued*)

Cox/ *QUANTITATIVE HEALTH RISK ANALYSIS METHODS*: *Modeling the Human Health Impacts of Antibiotics Used in Food Animals*

Ching & Ng/ *MARKOV CHAINS*: *Models, Algorithms and Applications*

Li & Sun/ *NONLINEAR INTEGER PROGRAMMING*

Kaliszewski/ *SOFT COMPUTING FOR COMPLEX MULTIPLE CRITERIA DECISION MAKING*

***** *A list of the more recent publications in the series is at the front of the book* *****