Mathematical Modeling in the Social and Life Sciences

Michael Olinick

WILEY

Mathematical Modeling in the Social and Life Sciences

Michael Olinick
Middlebury College

WILEY

DEDICATION

To Judy who, without resort to mathematics, is a model wife, parent, and citizen

Publisher:	Laurie Rosatone
Acquisitions Editor:	David Dietz
Content Editor:	Jacqueline Sinacori
Editorial Assistant:	Michael O'Neal
Product Designer:	Tom Kulesa
Cover Designer:	Kenji Ngieng
Marketing Manager:	Melanie Kurkjian
Associate Production Manager:	Joyce Poh
Senior Production Editor:	Jolene Ling

This book was set by MPS Limited.

Founded in 1807, John Wiley & Sons, Inc. has been a valued source of knowledge and understanding for more than 200 years, helping people around the world meet their needs and fulfill their aspirations. Our company is built on a foundation of principles that include responsibility to the communities we serve and where we live and work. In 2008, we launched a Corporate Citizenship Initiative, a global effort to address the environmental, social, economic, and ethical challenges we face in our business. Among the issues we are addressing are carbon impact, paper specifications and procurement, ethical conduct within our business and among our vendors, and community and charitable support. For more information, please visit our website: www.wiley.com/go/citizenship.

Library of Congress Cataloging-in-Publication Data

Olinick, Michael.
Mathematical modeling in the social and life sciences / Michael Olinick, Middlebury College.
pages cm
Includes index.
ISBN 978-1-118-64269-6 (pbk.)
1. Social sciences—Mathematical models. 2. Life sciences—Mathematical models. I. Title.
H61.O483 2014
300.1'51—dc23

2013044538

Printed in the United States of America
10 9 8 7 6 5 4 3 2 1

Contents

Preface

The goal of this book is to encourage the teaching and learning of mathematical model building relatively early in the undergraduate program. The text introduces the student to a number of important mathematical topics and to a variety of models in the social sciences, life sciences, and humanities. Students with some mathematical maturity and a strong secondary school background will find many chapters quite accessible. A standard first year calculus course is sufficient background for the remaining chapters. While many of the models use differential equations or some elementary linear algebra, no previous experience with these topics is assumed. The text material will help students gain the necessary knowledge. Appendices on sets, matrices, systems of linear equations, and functions of two variables provide additional background material.

Particular problems in political science, ecology, biology, evolution, medicine, psychology, sociology, economics, finance, anthropology, criminal justice, epidemiology, philosophy, religion, opera, and hospital planning provide the motivation for the development of tools and techniques employed throughout applied mathematics. These include

Differential Equations	Discrete Dynamical Systems,
Axiomatics	Probability Theory
Regular Markov Chains	Absorbing Markov Chain
Matrix Algebra	Least Squares Fitting Of Data
Simulation	Theory Of Games

The curricula in many social science and life science disciplines are becoming increasingly infused with the development and analysis of formal models. Students in such majors (particularly integrated biology/mathematics and economics/mathematics programs) need an introduction to the mathematical ideas and techniques just listed. This text provides one way of gaining this exposure in a single course.

I selected models primarily from the behavioral sciences for a number of reasons:

1. They show the rich variety of disciplines to which mathematics is making important contributions;
2. Because such models require less technical background knowledge than more traditional models in physics, chemistry, and engineering, students can examine in depth many different applications in a one-semester course;

3. Most students feel more familiar with social phenomena than with physical ones. They are more eager to challenge the assumptions of models and to develop alternative ones on their own;
4. These models provide a unique opportunity for a student with a minimal background in calculus to learn about some mathematical developments of the past century.

Structure of the Book

The first chapter introduces the idea of a mathematical model by reexamining a familiar physical example: what happens when an object falls toward the earth. The model is a classic one from elementary calculus that is a good exemplar of a continuous dynamic system. We also model, using a discrete dynamic approach, the important personal finance problem of controlling credit card balances. To demonstrate a theme that the same mathematical model can be used to investigate real world problems that seem on the surface to be different, we show how the credit card balance model may also be used to model population growth of nations. There is a discussion of the classification of models into deterministic, probabilistic, and axiomatic categories.

Chapters 2 through 5 concentrate on deterministic models. "Stable and Unstable Arms Races" (Chapter 2) presents L. F. Richardson's theories about the outbreak of war. The model is a linked system of two linear differential equations with constant coefficients. The mathematical analysis is kept to an elementary level; it exploits the idea that a derivative gives a good approximation to the behavior of a function near a point of tangency. I have presented this material several times to students during their first calculus course. We also indicate how ideas from linear algebra lead to an explicit solution of the system. Such systems appear in many different fields (eg. biology, ecology, environmental economics, pharmacokinetics) that employ compartment models. We also discuss numerical approximations to such systems via Euler's method.

As in subsequent units, Chapter 2 begins with a verbal description of a real-world problem and then proceeds to show how a mathematical model can be built that reflects the important assumptions. A good portion of each chapter is devoted to a mathematical analysis of the model, in which new mathematical tools are developed. After the analysis, we proceed to discuss how the model can be tested against real-world data and indicate how one might refine and improve the model.

Chapters 3 and 4 on ecological models go more deeply into the use of differential equations as modeling tools. First, models of population growth for a single species are introduced. These use the standard types of first-order differential equations leading to exponential and logistic growth. We show how to fit a logistic model to real world data. We also discuss discrete analogues to these models, illustrating how the discrete logistic model may exhibit chaotic behavior. In Chapter 4, nonlinear systems of differential equations provide the language for examining simple models of the fluctuations of populations of interacting species. In examining nonlinear systems, we emphasize the determination of stable points and characterizing their behavior through the use of approximation by linear models.

In Chapter 5, we investigate some discrete and continuous models of tumor growth with a focus on some very recent results about the dynamics of colorectal cancer tumors. We compare and contrast the logistic model presented earlier with the von Bertalanffy and

Gompertz models. This chapter also continues study of fitting data to a model and comparing different models' predictions to experimental data by the method of least squares.

Chapters 3–5, together with Chapter 13 on epidemics, are the only ones in the text which demand technical mastery of a year's study of calculus. The background material on differential equations and functions of two variables is presented in appendices.

Chapters 6 through 9 focus on axiomatic models. In the sixth chapter, "Social Choice and Voting Procedures," we discuss some of the injustices associated with commonly used voting mechanisms. The problem of interest becomes: "Can one construct a voting procedure which avoids these shortcomings?" We state and prove Arrow's Theorem that a seemingly plausible list of properties such a mechanism should satisfy turns out to be inconsistent. We also present some more recent theorems (with proofs) considering other "reasonable" but inconsistent axioms. The attractive features and interesting paradoxes associated with Instant Runoff Voting and Approval Voting are also treated in some depth. Chapter 6 concludes with an introduction to topological choice theory.

Chapters 7 and 8 present axiomatic treatments of some basic questions of contemporary measurement and utility theory. The existence of equilibrium prices in an exchange economy is the focus of Chapter 9 where we show how a Nash-Debreu approach using fixed point theory provides a critical insight. This chapter begins with a classic model of a 2-person economy using the ideas of an Edgeworth box, indifference curves, bargaining space, and Pareto solutions before moving to the more general approach. Beyond the usual demand for "mathematical maturity," there are no specific mathematical prerequisites for understanding these chapters 6–9. We make use of Brouwer's Fixed Point Theorem here and again in Chapter 16 on game theory. We do not include a full proof of Brouwer's Theorem but show its connection with the geometrically more plausible No Retraction Theorem.

Chapters 10 through 13 develop an extensive treatment of probabilistic models. There is particular emphasis on Markov processes because of their widespread use as models in the mathematical social sciences. The treatment is self-contained; no prior knowledge of probability or linear algebra is assumed. It has been my experience that much of the material in Chapters 10 and 11 can be assigned for self-study by the students. I present outlines of proofs for the main results about regular and absorbing Markov chains. The results are easily understood and applied by students with three years of high-school mathematics. I would reserve discussion of the proofs for a class which had already completed two or more years of college mathematics.

Once the background in probability theory is presented, there are a number of applications: population growth models, cultural stability (Chapter 12), paired-associate learning (Chapter 13), sports competition, and the spread of epidemics. The stochastic version of a simple deterministic model of population growth discussed in Chapter 3 is presented along with a comparison of the results obtained by the two approaches. Chapter 14 on epidemics also compares and contrasts deterministic and probabilistic models of the same problem. We also show how some of the ideas of modeling the spread of infectious diseases may be used to model the dynamics of rumors, the persistence of urban legends, and the control of binge drinking. A chapter available online shows an application of Markov processes to measuring recidivism in the criminal justice system.

Chapter 15 introduces computer simulation by examining the way in which a St. Louis hospital staff decided how many additional surgical and recovery rooms would be needed if it added a fixed number of new beds. We delve into the topic of agent-based simulation models in an additional chapter available online.

Chapter 16 discusses game theory, the first mathematical discipline specifically created to model human behavior. We begin with two difficult decisions facing Abraham in the book of *Genesis* and Tosca in Puccini's opera. The major classifications of games is introduced; one-person games (decision theory) is illustrated by David's decision to fight Goliath. Minimax mixed strategies in two-person zero-sum are presented. We show how to compute optimal mixed strategies for 2×2 games. We then illustrate how two-person nonzero-sum game theory sheds new insights into some famous stories of human sacrifice from the Old Testament and illustrate how the Prisoner's Dilemma game is embedded in *Tosca.* There is an in depth presentation of Nash equilibria including a proof of his famous existence theorem. An additional online chapter introduces Evolutionary Game Theory, a powerful new tool that is of increasing interest to biologists and economists. The final online chapter discusses agent-based models.

Five appendices provide background information on sets, matrices, systems of linear equations, functions of several variables, and differential equations.

A chart at the end of the preface shows the dependencies of each chapter on the earlier ones. There is more material in the text than can be covered in a one-semester course, so the chart will enable instructors to create a variety of different courses to emphasize their students' interests and mathematical preparation.

In my twelve-week course, I like to expose students to deterministic, axiomatic, and probabilistic models. In a typical term, we would cover the material in Chapters 1–4, 6, 10–13, 15 and 16. For a one-quarter (10 week) course, I would likely omit Chapters 13 and 15 and shorten the treatment in Chapter 4. The luxury of a full semester (15 weeks) would enable me to add two or three additional chapters.

An instructor wishing to emphasize models in the biological sciences might include the first five chapters followed by Chapters 10, 14, 15 and 18 . The core of a course directed toward students in economics could include Chapters 1, 2, 6–9, 10 and 16. These chapters incidentally provide deeper understanding of the mathematics used in the seminal work of at least three Nobel Prize winners in economics: Kenneth Arrow, Gérard Debreu, and John Nash.

To make a modeling course accessible to students at the earliest point in their undergraduate curriculum, experience in computer programming is not a prerequisite for the book. I believe the text is sufficiently flexible in its structure, however, to permit instructors to emphasize programming if they wish. Implementation of the discrete dynamical systems examples and simulations of the continuous systems can easily be done by students with access to many standard software packages such as Excel, STELLA, Maple, or Mathematica.

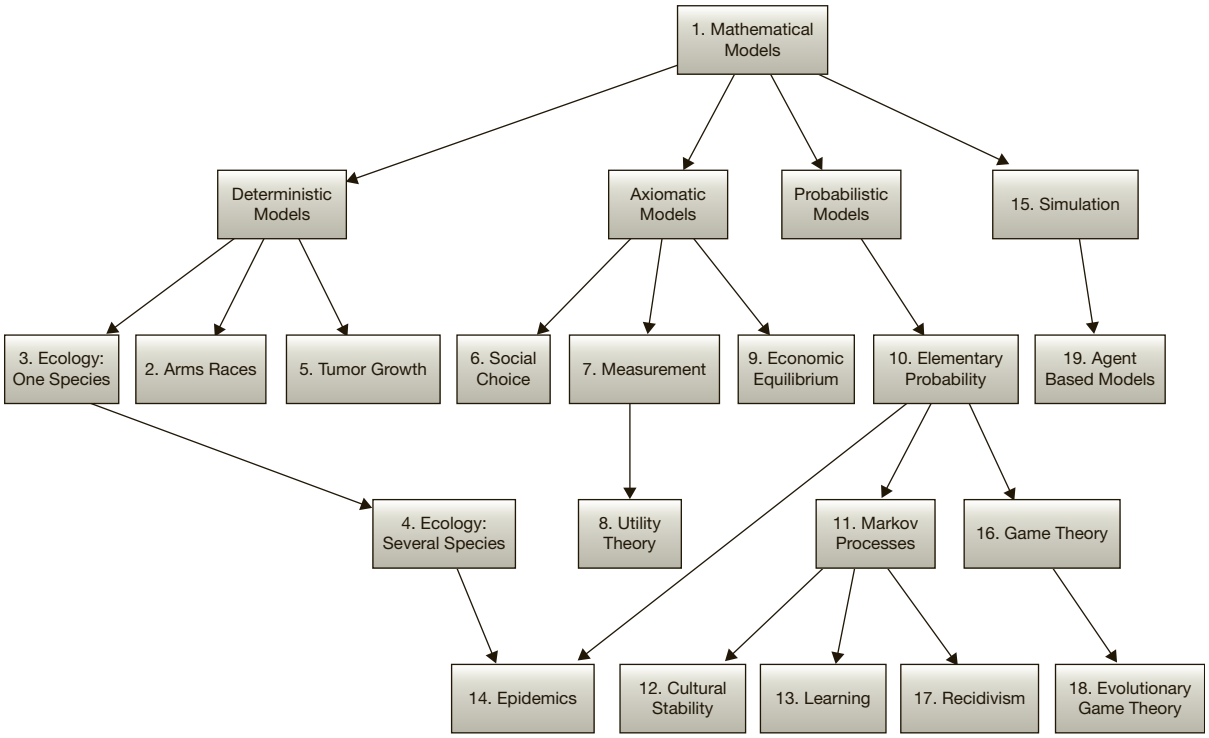
Concluding Remarks

In writing this book, I followed my belief that students can learn more about building mathematical models by studying critically, and in some depth, a relatively few models than they can by learning, in isolation, a large assortment of techniques. All the models presented are simple ones, in the sense that researchers have constructed more sophisticated and more realistic ones to model the same phenomena. Thus, readers will not find the latest developments in mathematical learning theory, for example, in this text, nor will they see a survey of the models commonly used by ecologists today in their study of interacting

populations. The models which are analyzed here have been for the most part, however, significant in the development of a mathematical approach to one or more disciplines.

The text hopefully encourages readers to go beyond the stage of examining the works of others and to begin to function as a model builder on their own. Many of the more than 700 exercises require the reader to create "minimodels" to solve the problems. At the end of each chapter there are suggested projects that demand the creation of new models or involve extensions of old ones. On the book's website, www.wiley.com/college/olinick there are extended lists of references for additional reading or further exploration.

The reader will also note that extensive space has been set aside for historical and biographical notes on the development of the mathematical models and the men and women who invented them. My hope is that these will dispel any lingering notions that mathematics is the creation of colorless automata. As in every other creative activity, mathematics and mathematical applications come from the minds of active individuals responding to the crucial social and cultural issues of their day.



Acknowledgements

I am very pleased to express my appreciation and gratitude to the many people who motivated the writing of this book and assisted in its development and production.

My interest in mathematical model building traces back to undergraduate years at the University of Michigan. There I was fortunate to be able to study mathematical psychology with Clyde Coombs and general systems with Kenneth Boulding. Another strong Ann Arbor influence was Anatol Rapoport whose book *Fights, Games and Debates* first led me to attempt a classroom presentation of Richardson's arms race model. Professor Rapoport carefully reviewed the entire manuscript of a very early draft of this book. His many valuable comments, criticisms, and suggestions resulted in a number of important revisions.

Grants over the years from the National Science Foundation, National Endowment for the Humanities, Alfred P. Sloan Foundation New Liberal Arts Program, Howard Hughes Medical Institute, and Ada Howe Kent Fund provided opportunities to learn and to write. Middlebury College generously gave me many resources to develop and finish this project, including academic leaves, travel support, materials, and support for student research assistants. My mathematics colleagues at Middlebury over the past four decades have sustained a most wonderful environment in which to teach and do mathematics.

Acknowledgments for many other helpful remarks are due to the individuals who examined part or all of the text at various stages of its development. They include Kenneth J. Arrow (Harvard University), Walter F. Bodmer (University of Oxford), Graciela Chilchilinksy (Columbia University), Gordon H. Bower and George B. Dantzig (Stanford University), Gerard Debreu (University of California, Berkeley), Hans Hoffmann (State University of New York, Binghamton), Matthew D. Johnston (University of Oxford) N. K. Kwak (St. Louis University), William F. Lucas (Cornell University), Ronald Mickens (Clark Atlanta University), Oskar Morgenstern, Irwin W. Sandberg (Bell Laboratories), Homer H. Schmitz (St. Louis University), and Harvey A. Smith (Oakland University).

I wish also to thank several reviewers who have suggested improvements in the book: Bernard Brooks (Rochester Institute of Technology), Charles Hampton (Wooster College), Dennis Higgins (SUNY Oneonta), Patricia Kenschaft (Montclair State University), Joseph Malkevitch (York College of CUNY), Susann Mathews (Wright State University), Walter Meyer (Adelphi University), Ramanjit Sahi (Austin Peay State University), Monir Sharobeam (Stockton State College), Martha Siegel (Towson University), and Tom Tucker (Colgate University).

To the several hundred Middlebury students who struggled with heaps of unbound and photocopied drafts of these chapters I also owe a debt of gratitude. Their enthusiastic response to

the material was a continuing stimulus to me to revise and improve my presentations. Special thanks go to Max Bacharach for preparation of a detailed solutions manual.

I am very grateful to Shannon Corliss, David Dietz, Jolene Ling, and Michael O'Neal of the staff of John Wiley and Sons for their constant cooperation, assistance, and endless patience.

I wish I could write "The author assumes no responsibility for any errors or omissions that may appear in this book" or at least assure you there are no mistakes, but the former is improper academic behavior and the latter is assuredly false. Please send me (molinick@middlebury.edu) comments on errors. I will try to post corrections on the book's website, www.wiley.com /college/olinick.

Mathematical Models

CHAPTER 1

The status of a science is commonly measured by the degree to which it makes use of mathematics.

—S. S. Stevens

It is still an unending source of surprise to me to see how a few scribbles on a blackboard or on a sheet of paper could change the course of human affairs.

—Stanislaw Ulam

I. Mathematical Systems and Models

A. Mathematical Systems

Science studies the real world. In their role as scientists, human beings want to discover the laws that govern observed phenomena. When we better understand phenomena, then we may make valid predictions about future behavior. In a more active capacity, such understanding can lead to intelligent efforts to control phenomena, or at least influence them.

In this book, we will examine how we can use *mathematical systems* as tools to help achieve some of these aims. Although you will examine some examples from the physical sciences, most of our attention will be on problems of primary interest to social and life scientists, philosophers, and humanists.

A *mathematical system* consists of a collection of assertions from which we derive consequences by logical argument. We commonly call the assertions the *axioms* or *postulates* of the system. They always contain one or more primitive terms that are undefined and that hence have no meaning inside the mathematical system.

A familiar mathematical system is that of plane geometry. Two of the primitive terms in this system are "point" and "line." As examples of axioms in this system, we have

AXIOM 1 Given two distinct points, there is a unique line containing the points.

and

AXIOM 2 Given a line L and a point p not belonging to L, there is a unique line that contains p and that is parallel to L.

Not all the terms in an axiom are necessarily primitive. The concept "parallel," which occurs in the statement of Axiom 2, is not itself primitive, but may be defined using primitive terms. We say that two distinct lines are *parallel* if there is no point that belongs to both of them. In a similar fashion, we can express every axiom of plane geometry in terms of the primitive concepts of the system.

To be *mathematically interesting*, it is necessary only that the set of axioms of the system be consistent and be "rich" enough to imply a number of nontrivial consequences. It has been known for hundreds of years that the standard axioms of plane geometry form a consistent collection. Nowhere in the large number of theorems that are implied by this set of axioms will you find two results that contradict each other.

The usual axioms of plane geometry may be modified without losing consistency. If Axiom 2 is replaced by

AXIOM 2′ Given a line L and a point p not belonging to L, there is no line containing p that is parallel to L.

then the resulting system is still consistent. This remarkable result was the surprising conclusion of many attempts to show that Axiom 2 (known as Euclid's Fifth Postulate) was itself a consequence of the other axioms of plane geometry.

We are not going to focus on systems that are only mathematically interesting. Our concern is with *scientifically interesting* systems. The criteria to be met for this label are that the primitive terms should correspond to, or at least be idealizations of, objects that exist in the real world, and that the axioms should reflect our experiences of how these objects relate to each other.

The mathematical system of plane geometry is also a scientifically interesting one. In fact, the system was evolved over a long period of time to put together in an organized and coherent fashion the observations that people had made about certain features of the world in which they lived.

One product of this mathematical system has been a collection of highly useful theorems about the measurement of the areas of many different regions of the plane. Applications of these results are too numerous and familiar to be mentioned here.

Quite often a system developed primarily, or solely, because of its mathematical interest has turned out also to be of fundamental scientific interest. The physical theory of relativity created by Albert Einstein (1879−1955) makes use of a geometry (called Riemannian geometry) in which Axiom 2′ rather than Axiom 2 is true. (Further discussion of this point is contained in Chapter 15.)

The axioms of a mathematical system will usually consist of statements about the existence or uniqueness of certain sets of elements, existence of various relations on these sets, properties of these relations, and so forth. Logical argument, or *deduction*, is applied to the mathematical system to obtain a set of mathematical conclusions. These conclusions are theorems about the primitive terms that were not immediately evident from the statements of the axioms.

B. Mathematical Models

When a mathematical system is constructed in an attempt to study some phenomenon or situation in the real world, we usually call it a *mathematical model*. There are many ways to model parts of the real world, and mathematics is only one of them. A road map of Dallas is

an example of a different kind of model. The map is a model of the city. If a motorist understands the symbols that are used in the map, then much information about the city becomes available in a package small enough to carry around in one's pocket. The motorist can use the map, for example, to plan a route from Southern Methodist University to the corner of Amberwood Road and La Manga Drive.

The road map is one representation of many important features of the city. But it omits many other features that may be crucial. Most road maps do not contain sufficient information to tell a motorist what is the speediest route to take between two points in the city during the morning rush hour, for example. The map is also almost useless to a door-to-door encyclopedia salesperson who wishes to find neighborhoods whose social and economic characteristics indicate good selling opportunities. For this purpose, a different kind of model of the city is needed. (For other kinds of models, see the discussion in Chapter 15.)

All models, be they physical or mathematical, are attempts to represent *certain* aspects of reality. Any effort to include *all* aspects would overwhelm us with detail and would require a model as large as the original object. The Argentine writer Jorge Luis Borges (1899–1986) captures the ludicrous nature of such an attempt is his short story "On Exactitude in Science":

> *In that Empire, the Art of Cartography attained such Perfection that the map of a single Province occupied the entirety of a City, and the map of the Empire, the entirety of a Province. In time, those Unconscionable Maps no longer satisfied, and the Cartographers Guilds struck a Map of the Empire whose size was that of the Empire, and which coincided point for point with it. The following Generations, who were not so fond of the Study of Cartography as their Forebears had been, saw that that vast Map was Useless, and not without some Pitilessness was it, that they delivered it up to the Inclemencies of Sun and Winters. In the Deserts of the West, still today, there are Tattered Ruins of that Map, inhabited by Animals and Beggars; in all the Land there is no other Relic of the Disciplines of Geography.*

A mathematical model of a complex phenomenon or situation has many of the advantages and limitations of other types of models. We omit some factors in the situation and stress others. In constructing a mathematical system, modelers must keep in mind the type of information they wish to obtain from it.

The relatively simple schematic diagram of Fig. 1.1 illustrates the role that mathematical models play in science.

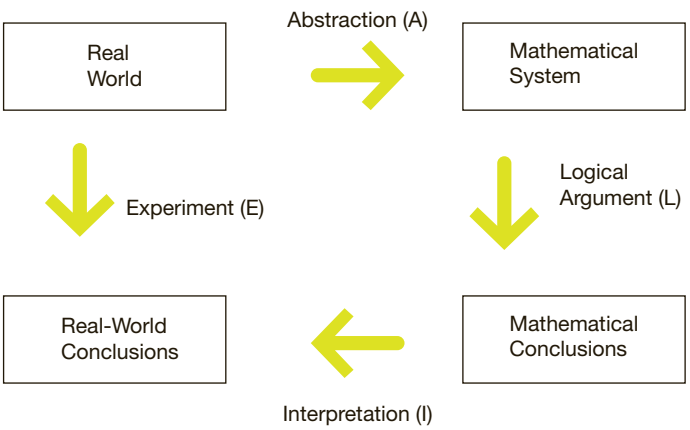


FIGURE 1.1 Schematic diagram of the modeling process.

Scientist begins with some observations about the real world. They wish to make some conclusions or predictions about the situation they have observed. One way to proceed, (E), is to conduct some experiments and record the results. The model builder follows a different path. First, she abstracts, or translates, some of the essential features of the real world into a mathematical system. Then by logical argument, (L), she derives some mathematical conclusions. These conclusions are then interpreted, (I), as predictions about the real world.

To be useful, the mathematical system should predict conclusions about the world that are actually observed when appropriate experiments are carried out. If the predictions from the model bear little resemblance to what actually occurs in the real world, then the model is not a good one. The modeler has not isolated the critical features of the situation being studied, or the axioms misrepresent the relations among these features. On the other hand, if there is good agreement between what is observed and what the model predicts, then there is some reason to believe that the mathematical system does indeed correctly capture important aspects of the real-world situation.

What happens quite frequently is that some of the predictions of a mathematical model agree quite closely with observed events, but other predictions do not. In such a case, we might hope to modify the model so as to improve its accuracy. The incorrect predictions may suggest ways of rethinking the assumptions of the mathematical system. One hopes not only that the revised model will preserve the correct predictions of the original one, but also that it will make further correct predictions. The incorrect inferences of the revised model will lead, in turn, to yet another version, more sophisticated than the earlier one. Thus, by stages, we develop a sequence of models, each more accurate than the previous ones.

We shall return to the general discussion of mathematical models and their advantages and limitations later in this chapter. To clarify some of the points that have already been made, it is useful to examine now a familiar mathematical model in some detail.

II. An Example: Modeling Free Fall

A. Formulation of the Model

Consider the fable that tells of Isaac Newton (1642−1727) sitting beneath the branches of an apple tree directly in the path of a descending apple. Whether or not Newton was ever actually struck by a plummeting piece of fruit, he was interested in the analysis of the motion of falling bodies. The real-world situation we wish to model here is described simply: an object, initially at some distance from the surface of the earth, is released; some time later, it strikes the earth.

This qualitative phenomenon is observed every day. If we are careful, we can measure the height of the object above the ground when it is released and also record the number of seconds that elapse before it strikes the ground. We wish to find a quantitative relationship between these observed values.

Our mathematical analysis of this situation begins by isolating the important concepts. Since we can measure distance and time, it is reasonable to develop a model in terms of these quantities. We will let t represent time in seconds and y represent distance above the ground in feet. As time varies, so does this distance. Thus, y is some function of time, $y = y(t)$, whose exact nature is as yet not known. We may start our stopwatch at $t = 0$ when the object is at distance y_0 feet above the ground.

Newton also realized that the mass m of the object was an important consideration in such a problem. One of the general laws of motion that Newton had formulated was that the product of the mass and acceleration of a moving body is equal to the sum of the forces acting on it.

For our first model, we will assume that there is only one force acting on the object, the gravitational attraction of the earth. Then Newton's Law of Motion has the familiar form

$$F = ma \tag{1}$$

where a is the acceleration of the body and F is the gravitational force.

Recall from elementary calculus that acceleration is the second derivative with respect to time of the position function $y(t)$, so that $ma = my''$. We also assume that the gravitational force F is proportional to the mass of the object with proportionality constant $g = -32\,\text{ft/sec/sec}$. Thus, $F = mg = -32m$. Substituting these assumptions into Eq. (1) produces

$$mg = my'' \tag{2}$$

or

$$y'' = -32 \tag{3}$$

B. Analysis of the Model

Eq. (3) is our mathematical model for a falling object. It is a simple second-order differential equation. We apply the tools of mathematical analysis (logical argument) to derive some mathematical conclusions. In this case, this means we should solve the differential equation. Integrate each side of Eq. (3) with respect to the variable t twice to obtain first

$$y' = -32t + C \tag{4}$$

and then

$$y = -16t^2 + Ct + D \tag{5}$$

where C and D are constants of integration. If we set $t = 0$ in Eq. (4), we find that C is equal to the value $y'(0)$, which we will denote by v_0.

Setting $t = 0$ in Eq. (5) gives a value of D equal to $y(0) = y_0$. Thus, we have

$$y' = -32t + v_0 \tag{6}$$

$$y = -16t^2 + v_0 t + y_0 \tag{7}$$

C. Interpretation of the Model

We may now interpret these mathematical conclusions as statements about the falling object. Since the derivative of the position function gives velocity, Eq. (6) is a prediction of

the velocity of the object at every instant, if v_0 is its initial velocity. In particular, if the object is simply released from a rest position, then $v_0 = 0$ and

$$y' = -32t \tag{8}$$

while

$$y = -16t^2 + y_0 \tag{9}$$

Eq. (9) can be used to answer our original question about the relation between the initial height of the object and the time it takes to reach the ground. When the object strikes the earth, we have $y = 0$. Substituting this fact into Eq. (9) gives the corresponding elapsed time, t_F, for the fall:

$$0 = -16t_F^2 + y_0 \tag{10}$$

or

$$t_F = \frac{\sqrt{y_0}}{4} \tag{11}$$

Our analysis thus gives a prediction for how long an object takes to fall a distance y_0 feet to the ground if it is released from rest. The analysis also yields a number of other predictions:

1. The velocity and position of the object at any time are independent of the object's mass. This follows because m is missing from Eqs. (3)–(11).
2. Using Eqs. (6) and (7), we can predict the velocity and position for situations in which the object is given any initial velocity. If v_0 is positive, then the model can be used to discuss what happens when the object is thrown upward, away from the earth, at the start of its motion.
3. If the object is released $(v_0 = 0)$ at height y_0, then the velocity of the object when it strikes the earth is $y'(t_F) = -32\, t_F = -32\left(\sqrt{y_0}/4\right) = -8\sqrt{y_0}\,\text{ft/sec}$.

D. Tests and Refinements of the Model

Let us concentrate, for a moment at least, on the predictions of Eq. (11). We can test the validity of our model by dropping objects from various heights, recording the time of fall, and comparing this number to the predicted value. We record some typical values in Table 1.1.

The first few values in the table seem reasonable and consistent with everyday experience, but can we say the same for the final entries? The last entry in Table 1.1 indicates that the predictions of a model can sometimes be shown to be incorrect without actually performing any physical or social experiments. The number 240,000 represents the approximate distance (in miles) between the earth and the moon. According to the model, an object leaving the surface of the moon should fall to the earth in about 2½ hours. In

Table 1.1

Initial Height (y_0)	*Predicted Time of Fall* ($t_F = \sqrt{y_0}/4$)
16 ft	1 sec
100 ft	2.5 sec
625 ft	6.25 sec
25,600 ft	40 sec
240,000 miles	2.5 hours

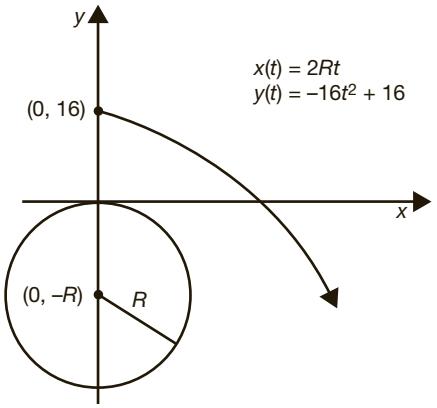


FIGURE 1.2 Motion in the plane. Whether the object hits or misses the disk depends on its initial position and velocity.

particular, if whoever is holding the moon in place should suddenly let go, this model asserts that the moon would crash into the earth about 150 minutes later. Since no one is really holding on to the moon, why hasn't it fallen?

This "thought experiment" indicates that the model cannot be accurate for all values of y_0. Where does the model go astray? We know, as did Newton, that, in the first place, the force of the earth's gravitational attraction on an object varies with the distance between the object and the *center* of the earth. The farther away the object is, the smaller is the attraction. It is only when the object's distance from the surface of the earth is small in comparison to the earth's radius (about 4,000 miles) that it is reasonable to treat the gravitational force as constant. To refine the model, the first correction is to replace the simple constant g by an appropriate decreasing function of y.

This refined model would still predict that the moon will eventually crash into the earth, although it will take somewhat longer than 2 or 3 hours. The fact that the moon has not done this indicates yet another difficulty with our model. The moon, or any object moving in three-dimensional space, has components of motion in three mutually perpendicular directions. The model considers only motion in one direction. Even though the force acts along that line of direction, it turns out that the object itself need not move exactly along that line.

For a simplified example, consider motion in the plane. Construct a normal Cartesian coordinate system with x- and y-axes and origin O. The position of an object at any time t is given by a pair of numbers $(x(t), y(t))$ representing the coordinates of its location as functions of time. Imagine a circular disk of radius R with center at $(0, -R)$ and suppose that the moving object under consideration is at the point (0, 16) at time $t = 0$; see Fig. 1.2. Assume that the only force acting on the object is a constant force of -32 in the vertical direction.

If the vertical component of velocity at time $t = 0$ is zero, then the y-coordinate is given, according to Eq. (9), by

$$y(t) = -16t^2 + 16 \tag{12}$$

If the horizontal component of velocity at time 0 is c ft/sec, then the x-coordinate of motion is

$$x(t) = ct \tag{13}$$

since there is no force acting in a horizontal direction.

Should the value of c be zero, the object will slide directly down the y-axis and will hit the disk at the origin at time $t = 1$. If c is nonzero, then the motion is more complicated. From Eq. (13), we have $t = x/c$ so that Eq. (12) may be rewritten as

$$y = -\frac{16x^2}{c^2} + 16 \tag{14}$$

Eq. (14) indicates that the path of motion in the (x, y)-plane will be a parabola. See Fig. 1.2.

In particular, if c should be equal to $2R$, then the object will never hit the disk! For during the first second of motion, $0 < t < 1$, the object is moving in the first quadrant, since both x- and y-coordinates are positive. At time $t = 1$, the object is at the point $(2R, 0)$. For $t > 1$, the y-coordinate is negative, while the x-coordinate is larger than $2R$. Since no point on the disk has an x-coordinate greater than R, the parabola will not intersect the disk.

As we have just seen, the initial horizontal speed of the moving object must be considered before determining whether or not the object will hit the disk. The initial position of the object also must be examined. For if $c = 2R$, but the object is at $(0, 4 - R)$ at time 0, it may hit the disk. For example, when $t = 1/2$, we would have $x = 2R(1/2) = R$, while $y = -16(1/2)^2 + (4 - R) = -R$. The point $(R, -R)$ lies both on the disk and on the parabolic path of the object.

A similar but more complicated analysis is possible for an object moving in three-dimensional space under the influence of the earth's gravitational attraction. Here the force is directed along a line between the object and the center of the earth. Depending on the object's initial distance from the earth and the various components of its initial velocity, it will either crash into the earth or orbit about it in an elliptical path. [See Chapter 3 of Simmons 1991 for a detailed derivation.]*

This more complex mathematical model may be forced upon us if we are planning a trip to the moon or if we must solve some other serious astronomical problem. The simple model of Eq. (3) breaks down for objects relatively far from the earth and for objects moving with great speed. If our concern is for apples falling from trees, or for other situations in which a relatively small object begins its fall from rest from a position fairly close to the surface of the earth, is the simple model still good?

* You will find references in brackets on the text's website, www.wiley.com/college/olinick

Eq. (11) was the main prediction of the simple model. We have already indicated how we can test the validity of this equation by dropping objects from various heights and timing the duration of their fall. There is an even simpler experiment to test the prediction that the elapsed time t_F is independent of mass: simultaneously release two objects of moderate but different masses from the same height and observe whether they reach the ground together.

This is the type of experiment allegedly conducted by Galileo (1564–1642), who was the first person to derive the equations leading to Eq. (11). Legend has it that Galileo dropped balls of different weights from the top of the Tower of Pisa and timed their descent. Although there appears to be as much truth in this tale as in the story of Newton and the apple (it has been conjectured that Newton made up this tale in response to repeated inquiries from those seeking a simple explanation for his deep discoveries), Galileo did conduct many experiments with objects rolling down inclined planes; interesting discoveries about his experimental and theoretical work are still being made [see Drake 1973 and 1975].

In any case, the experiment we have described works out fairly well in practice and is a standard laboratory assignment in many introductory physics courses. The experiment does not always produce the desired results, however. Once, I dropped a crumpled sheet of paper out the window of my ninth floor office at the same time my 2-year-old son released a sheet of paper that had been folded into the shape of a glider. The two sheets of paper came from the same pad, so their masses were essentially the same. According to our model, they certainly should have reached the ground at about the same time. My crumpled wad plummeted straight to the ground in a matter of seconds while the glider actually rose several feet before gradually floating to the earth a few minutes later. Why has our model failed us again?

The answer is easy. We have neglected in the model some important forces that act on our falling object: air resistance and wind currents. Recall that Newton's Law really asserts that the *sum* of the forces acting on an object equals the product of mass and acceleration:

$$\sum_i F_j = ma \tag{15}$$

To refine our model to make it more realistic, we have to account for these other forces in our differential equations. There are some relatively easy ways to include air resistance in this model (see Exercise 26), but the representation of wind currents can be a very tricky mathematical problem.

The moral of this story is that if you want a model that gives realistic predictions over a broad range of relevant variables (in this case, distance, mass, density, initial velocity, and so on), you must be willing to deal with complex mathematical systems. If you seek a simple and elegant model, you must be careful to describe its somewhat limited applicability. Thus, if Eq. (3) is to be an accurate model of a falling object, we must restrict ourselves to situations in which the object is of moderate mass, is dropped relatively close to the earth's surface, and falls through a vacuum. If we design an experiment by building a tube and then pumping out the air before dropping the object inside it, the observed data will be quite close to those predicted by Eq. (11). On the moon, where there is no atmosphere, all objects dropped from the same height should fall to the surface in the same amount of time; in the absence of an atmosphere all objects fall at the same rate; Apollo 15 astronaut David Scott demonstrated this conclusion by dropping a feather and a hammer on the moon's surface. [You can find a link to a video of this experiment at http://nssdc.gsfc.nasa.gov/planetary/lunar/apollo_15_feather_drop.html.]

III. Discrete Examples: Credit Cards and Populations

The variables in the model of free fall we have just examined (time, vertical position, and velocity) are each *continuous*. The independent variable, time, takes on all real values between the beginning and end of some interval. As time varies, the height of the object above the ground and its speed also change in a smooth manner. Indeed, the fact that position and velocity are differentiable functions of time is what makes it possible for us to apply the tools of calculus to analyze the model.

Many real-world phenomena, however, do not change in a continuous manner. The number of members in your family cannot smoothly increase from three to four, for example, taking on all the intermediate values. Family size remains constant for relatively long periods of time and then suddenly jumps up by one or more (with a birth or multiple births) or diminishes abruptly (with a death). The amount of money in your savings account jumps from one level to another as deposits, withdrawals, and interest payments are made. The enrollment at your college or university is always a whole number of students and hence can't change by a fraction. Other examples are the number of automobiles produced each year, the hourly wages of a fast-food worker, the inventory of nuclear weapons in a nation's arsenal, the number of countries in the United Nations, and so forth.

Quantities that cannot take on all intermediate values between two levels they can achieve or that cannot change at every possible instant but only at prescribed moments are called *discrete* variables. Discrete mathematics, the study of the tools to deal with discrete variables, is a rich and rapidly developing discipline. Much of the recent focus has been on *discrete dynamical systems*: the investigation of quantities that change only at discrete points in time. Simple discrete dynamical systems may produce very complex and chaotic behavior.

As an initial example of a discrete dynamical system, let us consider the balance in a personal credit card account, the total amount you owe to the bank that issued the card. Suppose that the initial balance B_0 is \$1,000 and that you make a monthly payment of $p = \$10$ to reduce the balance. Then the balance after 1 month is

$$B_1 = B_0 - p = \$1000 - \$10 = \$990$$

And the amount after 2 months would be B_2, where

$$B_2 = B_1 - p = \$990 - \$10 = \$980.$$

We may also write B_2 as

$$B_2 = B_1 - p = (B_0 - p) - p = B_0 - 2p.$$

It is easy to see that the balance B_n after n months would be

$$B_n = B_{n-1} - p \tag{16}$$

or

$$B_n = B_0 - np \tag{17}$$

and hence, the number n of months necessary to pay off the balance $(B_n = 0)$ would be

$$n = \frac{B_0}{p}$$

For our very simple example, it takes 1000/10 = 100 months to reduce the current balance to zero.

This simple model of the dynamics of a credit card balance ignores one extremely important real-world fact: banks charge interest on the balance that remains to be paid each month. Once a month, the bank adds a "finance charge" to the existing balance to create a new balance. The finance charge is a fixed interest rate percentage multiplied by the outstanding balance. Suppose, for example, that the interest rate r is 1.5% per month. Then the new balance would be given by

$$B_{new} = B_{old} + rB_{old} = (1 + r)B_{old} = (1.015)B_{old} \tag{18}$$

If you made no monthly payment at all, then

$$B_n = (1 + r)B_{n-1} \tag{19}$$

so that

$$\begin{aligned} B_1 &= (1 + r)B_0 \\ B_2 &= (1 + r)B_1 = (1 + r)^2 B_0 \\ B_3 &= (1 + r)B_2 = (1 + r)^3 B_0 \end{aligned}$$

and, in general,

$$B_n = (1 + r)^n B_0 \tag{20}$$

Table 1.2 shows the how an original balance of $1,000 grows over a 20-month period with a 1.5% monthly interest charge.

To make our model more realistic, suppose the bank charges $r = 1.5\%$ per month on the unpaid balance and you also make a monthly payment $p = \$10$. Then the new balance is calculated from the old by

$$B_{new} = B_{old} + r\,B_{old} - p = (1 + r)B_{old} - p = s\,B_{old} - p \tag{21}$$

where $s = 1 + r$. Eq. (21) is sufficient for a spreadsheet program to compute the balance for a sequence of months if it is given the initial balance, interest rate, and monthly payment. A careful study of the consequences of Eq. (21) enables us to find a direct way to calculate the balance at the end of any particular month without deriving the balance for all the previous months.

Table 1.2

n	B_n	n	B_n
0	\$1,000.00	11	\$1,177.95
1	\$1,015.00	12	\$1,195.62
2	\$1,030.23	13	\$1,213.55
3	\$1,045.68	14	\$1,231.76
4	\$1,061.36	15	\$1,250.23
5	\$1,077.28	16	\$1,268.99
6	\$1,093.44	17	\$1,288.02
7	\$1,109.84	18	\$1,307.34
8	\$1,126.49	19	\$1,326.95
9	\$1,143.39	20	\$1,177.95
10	\$1,160.54		

Note that from Eq. (21), we have

$$B_1 = sB_0 - p$$
$$B_2 = sB_1 - p = s(sB_0 - p) - p = s^2 B_0 - sp - p = s^2 B_0 - p(1+s)$$
$$B_3 = sB_2 - p = s(s^2 B_0 - p(1+s)) - p$$

or

$$B_3 = s^3 B_0 - p(1 + s + s^2)$$

Similarly,

$$B_4 = sB_3 - p = s\,(s^3 B_0 - p(1 + s + s^2)) - p$$

or

$$B_4 = s^4 B_0 - p(1 + s + s^2 + s^3)$$

In general, we would have

$$B_n = s^n B_0 - p(1 + s + s^2 + s^3 + \cdots + s^{n-1}) \tag{22}$$

This last formula can be simplified.

$$\text{Let } T = (1 + s + s^2 + s^3 + \cdots + s^{n-2} + s^{n-1}) \tag{23}$$

Then

$$\begin{aligned} sT &= s\left(1 + s + s^2 + s^3 + \cdots + s^{n-2} + s^{n-1}\right) \\ &= s + s^2 + s^3 + \cdots + s^{n-1} + s^n \end{aligned} \tag{24}$$

Subtracting Eq. (24) from Eq. (23) gives

$$T - sT = \left(1 + s + s^2 + s^3 + \cdots + s^{n-2} + s^{n-1}\right) - \left(s + s^2 + s^3 + \cdots + s^{n-1} + s^n\right)$$

Thus,

$$(1 - s)\,T = 1 - s^n$$

or

$$T = \frac{1 - s^n}{1 - s} = \frac{1 - (1 + r)^n}{1 - (1 + r)} = \frac{1 - (1 + r)^n}{-r} = \frac{(1 + r)^n - 1}{r}$$

and hence,

$$B_n = s^n B_0 - p\left(1 + s + s^2 + s^3 + \cdots + s^{n-1}\right) = s^n B_0 - pT$$

so,

$$B_n = (1 + r)^n B_0 - p\frac{(1 + r)^n}{r} = (1 + r)^n \left(B_0 - \frac{p}{r}\right) + \frac{p}{r} \tag{25}$$

Table 1.3 shows the balance for each of the first 15 months on a credit balance with an initial charge of \$1,000, a monthly 1.5% interest rate, and monthly payments of \$10 and \$25. Note that the balance owed the bank keeps increasing if the monthly payments are as small as \$10, but that the balance will decrease when payments are increased to \$25.

If we extend the calculations beyond 15 months, we find that a \$25 monthly payment will reduce to balance to \$13.42 after 61 months. Thus, 61 payments of \$25.00 plus a single payment of \$13.42 will pay back the loan. Note that you will have paid the bank a total of \$1,538.42 for the initial charge of \$1,000.

The mathematical model represented by Eq. (21) governs many other situations when money is borrowed to enable a consumer to make a large purchase. Examples include automobile loans, college tuition loans, or mortgages to finance buying a home. For a home loan, the amount borrowed may exceed several hundred thousand dollars and the payback period usually extends 20 or 25 years.

As an example of home mortgage situation, suppose you borrow \$150,000 with an annual interest rate of 8.4% and monthly payments of \$1,300. The balance after 236 months would then be \$447.73. Under this payment plan, you would return to the bank \$307,247.73, so the total interest amount is \$157,247, 73. In this example, the total interest paid exceeds the amount originally borrowed.

Table 1.3

Month	$p = \$10$	$p = \$25$
0	\$1,000.00	\$1,000.00
1	\$1,005.00	\$990.00
2	\$1,010.08	\$984.85
3	\$1,015.23	\$979.62
4	\$1,020.45	\$974.32
5	\$1,025.76	\$968.93
6	\$1,031.15	\$963.47
7	\$1,036.61	\$957.92
8	\$1,042.16	\$952.29
9	\$1,047.80	\$946.57
10	\$1,053.51	\$940.77
11	\$1,059.32	\$934.88
12	\$1,065.21	\$928.90
13	\$1,071.18	\$922.84
14	\$1,077.25	\$916.68
15	\$1,083.41	\$910.43

Assuming the amount of the loan and the interest rate is fixed, borrowers are often concerned with the related questions of the size of the monthly payment and the number of payments they will have to make to reduce the outstanding balance to 0.

Eq. (25) tells us that the balance will be 0 when

$$(1+r)^n\left(B_0 - \frac{p}{r}\right) + \frac{p}{r} = 0 \tag{26}$$

If we solve Eq. (26) for p, we find that to pay off a loan of B_0 at a monthly interest rate r in n months requires a monthly payment p where p is given by

$$p = \frac{rB_0(1+r)^r}{(1+r)^n - 1} \tag{27}$$

To see how the size of the regular payments affects the number of months n to pay off the loan, we solve Eq. (26) for n:

$$n = \frac{\ln\left(\frac{p}{p - rB_0}\right)}{\ln(1+r)} \tag{28}$$

It's instructive to examine a graph of n as a function of p, for a typical home loan. With $Bo = \$150{,}000$ and $r = .07$, the graph is shown in Fig. 1.3.

Observe from this graph that a modest increase in the monthly payment can dramatically reduce the length of time it takes to repay the loan and the total amount of interest

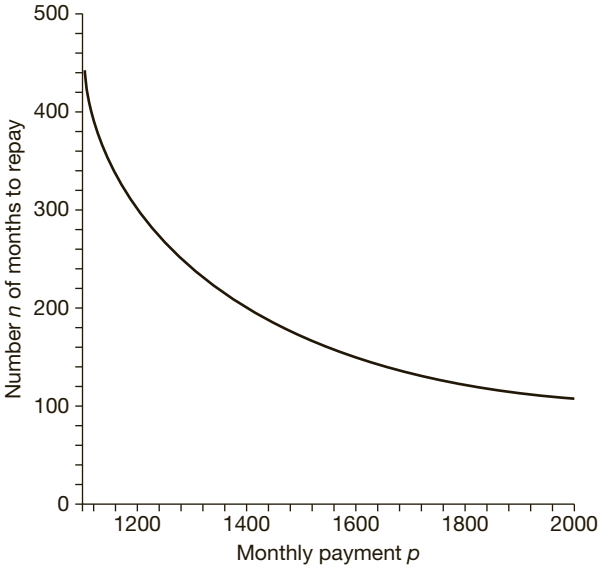


FIGURE 1.3 Decline in number of months to pay off loan as size of monthly payment increases.

Table 1.4

Year	Predicted	Actual	Error	Relative error
1960	179.323	179.323	0	0%
1970	202.770593	203.302	0.531407	0.26%
1980	227.515572	226.546	−0.969572	−0.43%
1990	253.629721	248.71	−4.919721	1.98%
2000	281.1888	281.421	0.2322	0.083%
2010	310.273	308.746	1.52	−0.495%

you give to the bank. A monthly payment of $1,200, for example, requires nearly 25 years (298 months) to repay. The total paid to the bank is about $357,600, of which $207,600 is interest. If you can increase the regular payments by $200 so that you give the bank $1,400 at end of each month, then the mortgage will be paid off in about 16.5 years (198 months). You will have paid the bank about $277,200, an overall savings of more than $80,000 compared to the $1,200 payments.

Similar reasoning can be used to model the growth of the U.S. population during the last half century. A simple model posits that the annual "internal" growth rate, the difference between the birth and death rates was $r = .0054$ percent and that there was a net migration (immigration − emigration) of $p = 1.32$ million people per year. Thus, with an initial population of P_0 million, the population after n years would be given by

$$P_n = (1 + r)\, P_{n-1} + p$$

Table 1.4 shows a comparison of the model's predictions and Census Bureau data. This model predicted the 2010 population to be 310.273 million and estimates the 2020 population at 340.966 million. In December 2010 the Census Bureau reported the

Table 1.5

Year	Predicted	Actual	Error	Relative Error	% Error
1960	179.323	179.32	0	0	0
1950	157.10477	150.216	6.88877	0.045859	−4.59%
1940	136.051447	131.669	4.382447	0.033283	−3.33%
1930	116.101955	122.775	6.673045	0.054352	5.44%
1920	97.1984194	105.711	8.5125806	0.080527	8.05%

U.S. population to be 308.746 million. The model's prediction was off by less than 0.5 percent. Census experts predict the 2020 population to be 341.387 million.

While we can only wait to see how accurate this model will prove to be in the future, we can also test out its reasonableness by running it backward to see what the U.S. population would have been in previous decades if its dynamics had been governed by the same equation.

$$P_{new} = (1 + r)\, P_{old} + p$$
$$P_{old} = \frac{P_{new} - p}{1 + r}$$

Table 1.5 compares the actual census data with the model's predictions.

IV. Classification of Mathematical Models

The simple mathematical model of a falling object given by the equation $y'' = -32$ and the possible refinements of it that you have seen in the preceding sections are examples of what are called *deterministic models*. The assumption behind a deterministic model is that the entire future behavior of the system is exactly and explicitly determined by the present status of the system and the forces acting on it. In other words, if we know everything about the system at a particular moment (the state of each of its component variables and the forces impinging on them), then we can predict its behavior at every future instant. This was the belief that led to the very fruitful development of the physical sciences. Much of our understanding of the behavior of physical systems comes from deterministic models employing the tools of calculus. Powerful analytic techniques were developed to analyze more and more complicated models. The availability of these techniques and the predictive successes of these models in the physical sciences motivated many thinkers to employ similar models in the study of social and biological systems. Chapters 2−5 of this text explore some of these deterministic models in detail.

There are several important objections to the use of such deterministic models in the social and life sciences. In the first place, some philosophers have argued that deterministic models of social systems must necessarily assume that human beings have no free will; few people are willing to accept this view of humans.

A second objection arises from the discovery of Werner Heisenberg (1901−1976) in the early part of the 20th century that purely deterministic models are insufficient even to study physical processes. Heisenberg showed that it is impossible, even in theory, to know the exact state of a physical system: the act of observation itself changes the system. We can

conclude from deterministic models only the *average* behavior of a group of atoms, but we cannot assert with certainty anything about the future course of an individual atom. This observation has fundamentally affected the physical sciences (see the discussion at the beginning of Chapter 10) and led to the introduction of *probabilistic models*.

Probabilistic models are also predictive models. The basic assumption of such models is that the system under investigation can occupy one of several different possible states at each moment, with different probabilities. If we know the probability distribution governing the system at the present moment and the forces acting upon the system, then we can predict the probability distribution at subsequent times. Thus, a probabilistic model of an object falling to the surface of the earth will predict for each time t_0 after the object is released the probability, or likelihood, that the object has reached the ground by time t_0. A substantial part of this text, centered around Chapters 10−14, discusses probabilistic models.

Since deterministic models often provide good approximate predictions and since they usually employ the familiar tools of calculus and differential equations, they are still widely used in the physical as well as social and life sciences. In Chapter 10 and again in Chapters 11 and 14, you will see the differences between a deterministic and a probabilistic model of the same situation.

A third objection to the calculus-oriented deterministic model centers on the use of the calculus. Since calculus was largely invented to help solve physical problems, there is no intrinsic reason why it should be the appropriate tool for the formulation and investigation of all social and biological systems, even if a deterministic approach is assumed. Indeed, new mathematical tools such as the theory of games, linear programming, and graph theory have been forged in recent decades to analyze such systems. Chapters 16 and 18 focus on the theory of games.

The deterministic and probabilistic approaches we have been discussing share the property of being *predictive* in nature; they both aim at saying something about the future (or perhaps past) of a system whose present state is fairly well described. These can be contrasted with models that are primarily *descriptive* in nature. Descriptive, or *axiomatic models*, as they are sometimes called, are highlighted in this text in Chapters 6−9.

The model in Chapter 6 is concerned with the possible existence of a voting mechanism that is constrained to satisfy certain "fairness" restrictions. Chapters 7 and 8 also present descriptive models; these describe different types of measurement and utility and when each can be used. In Chapter 9, we examine the existence of a set of prices in an economy that will guarantee there is sufficient supply to goods and services to meet the total demand.

The development of such axiomatic models hopefully will broaden your view of mathematics. Many people still believe that "mathematics is the study of numerical and geometrical concepts." Such definitions were common in texts and dictionaries even in recent years. Mathematicians today have a much wider view of their discipline. As John Kemeny and J. Laurie Snell [1962] phrase it:

> *Mathematics is best viewed as the study of abstract relations in the broadest sense of that word. From this point of view it is not surprising that mathematics is applicable to any well-defined field. Whatever the nature of the phenomena studied in a given social science, their various components do bear certain relations to each other, and once one succeeds in formulating these abstractly and precisely, one is in a position to apply the full machinery of mathematical analysis.*

V. Uses and Limitations of Mathematical Models

The continuing development of a useful coordinated science of social, biological, and physical behavior is an important challenge for us. In what ways can we expect mathematics to help?

If we begin to analyze even a simple situation involving interpersonal relations, for example, the first thing we observe is how complex the situation really is and just how many different variables are present. It is difficult to cope with so many factors adequately in an intuitive and discursive way. As long as our formulation remains vague and imprecise, our observations are likely to be muddled, unclear, and poorly understood by others.

Casting our thoughts into a mathematical model will have immediate advantages. Mathematics is a precise and unambiguous language. In order to use it, we must first clarify to ourselves the underlying assumptions we are making. The mathematical model forces us to organize our thoughts in a more systematic way, and this should contribute to the clarity of our thinking.

Once the model has been formulated, it is possible to use mathematical tools to derive new observations or conclusions from the model that may have escaped us if we proceeded with a more intuitive approach. These conclusions will not only shed light on the assumptions we have originally made about the system under study, but they will also suggest further experiments and observations that will lead to more complete knowledge of the system.

Quite often, the modeler discovers that a mathematical formulation of a problem turns out to be the same as someone else's formulation of what appears to be a totally different situation. The logistic equation—presented in detail in Chapter 3—has been used, for example, to model the growth of populations, the spread of infectious diseases, rates of chemical reactions, and consumer demands for commercial goods. Thus, the use of mathematical models can reveal unsuspected relations among superficially disparate systems that have the same basic underlying structure.

As we have seen in the discussion of falling objects, a simple mathematical model cannot precisely mimic the behavior of a real-world phenomenon. Some aspects of the real world are highlighted, and others are neglected or perhaps ignored. Mathematics is but one tool to be used in gaining an understanding of the real world, and it must be supplemented by other approaches.

As you study the models presented in this text, you may wish to consult Fig. 1.1 from time to time. Our emphasis will be on the steps of abstraction, logical argument, and interpretation. It has often been stated that it is difficult, perhaps impossible, to write down rules for the abstraction process. Many model builders believe that doing so is an art, rather than a science, and that it is best learned through the careful study of selected examples and repeated practice on the part of the apprentice in constructing models. The text material, exercises, and projects in this book have been designed to provide you with such practice.

The step we have labeled "logical argument" is the one that is most familiar to mathematics teachers and their students, while the process of "interpretation" follows fairly easily from the original formulation of the model.

The final critical step in mathematical modeling is a comparison with the interpreted results of the model with the observations obtained from direct interaction with the real world. I have tried to show in this text a number of places where an originally simple model is refined in the light of these comparisons. The measurement of how closely the model fits the real world is, in general, a complicated problem that involves the full use of statistical techniques. We do not have space in this book to delve deeply into this field.

EXERCISES

I. Mathematical Systems and Models

1. Is "angle" a primitive term in geometry? What about "belonging to"? What are the other primitive terms?

2. To what extent are the axioms of geometry "self-evident truths" rather than expressions of cumulative experience?

3. A *realization* of a mathematical system is a physical representation of the set of axioms in the sense that real-world quantities can be found to take the place of the primitive terms in such a way that the statements of the axioms can be seen to be true relations among the real-world quantities. Is there such a real-world representation of the axioms of plane geometry? Show that a mathematical system that has a realization must also have a consistent set of axioms.

4. A *projective plane* P is a mathematical system consisting of primitive terms called "points" and "lines" and a relation of "containment" satisfying the three following axioms:

(A) Any two distinct points are contained in a unique line.

(B) Any two distinct lines contain a unique point.

(C) There are four points such that no three are contained in the same line.

(a) Show that there are four lines in P, no three of which contain the same point.

(b) Show that every line in P contains at least two points.

(c) Show that every line in P contains at least three points.

5. Find a realization of the projective plane with exactly seven points. Is there a realization with fewer points?

6. Find a realization of the projective plane with an infinite number of points. Is the projective plane a mathematically interesting system?

7. Consider the mathematical system S consisting of primitive terms "point" and "line" and a relation of "containment" satisfying the following four axioms:

(A) Each line contains a nonempty collection of points.

(B) Any two distinct lines contain a point.

(C) Each point is contained in exactly two lines.

(D) There are precisely four distinct lines.

(a) Find a realization of S.

(b) Prove that each line in S contains precisely three points.

(c) Show that there are precisely six distinct points in S.

(d) Is this a mathematically interesting system?

8. A set of axioms is *independent* if no axiom in the set can be derived logically from the others.

(a) Show that axiom A_1 is independent of Axioms A_2, A_3, . . . , A_n if the system consisting of A_2, . . . , A_n has a realization in which A_1 is false.

(b) Use (a) to formulate a criterion for the independence of a set of axioms.

9. In what sense is a "model airplane" a model of an airplane?

10. How do the literary concepts of "simile" and "metaphor" function as models?

11. List some other types of models you have encountered.

II. An Example: Modeling Free Fall

12. Suppose that the model of free fall is $y'' = g$ where g is an unknown constant.

(a) Analyze this model mathematically. What is the analogue of Eq. (11)?

(b) Describe an experiment that would determine the value of g.

13. Assuming that v_0 is nonzero, use Eq. (7) to find t_F. Comment on the fact that you obtain two different values for t_F. Show that one of these can be discarded if v_0 is negative. What happens if v_0 is positive?

14. If v_0 is nonzero, what is the velocity of the object when it hits the ground?

15. Suppose that a ball is thrown upward from a height of 3 feet with an initial velocity of 8 ft/sec. Use the model of Eq. (3) to analyze its motion. In particular, find

(a) The maximum height the ball reaches

(b) The time at which the ball is again 3 feet from the ground

(c) The number of seconds the ball is in the air

(d) The speed with which the ball strikes the ground

(e) The maximum speed the ball achieves

16. A man falls off the top of a building 1,024 ft high. Three seconds later, Wonder Woman (who can fly) arrives at the point from which the man fell. She dives down in an effort to save him. If she is capable of an initial velocity of 50 ft/sec, will she reach him before he hits the ground?

17. A more correct version of Newton's Law than (1) is that force is equal to the derivative, with respect to time, of mass times velocity.

(a) Show that if the mass is constant, then this law reduces to (1).

(b) Analyze the motion of a falling bucket of sand with a hole in it. The bucket originally t weighs 10 pounds but loses sand at a constant rate of t pound each second.

Problems 18−21 concern the mathematical model developed in connection with Fig. 1.2.

18. For what range of values of c will an object initially at (0, 16) eventually strike the disk?

19. If $c = 2R$, for what values of y_0 will an object initially at $(0, y_0)$ eventually strike the disk?

20. A motorcycle leaves the edge of a tall cliff with a velocity of 60 mph in a horizontal direction. Develop a mathematical model for the subsequent motion.

21. An airplane releases a nuclear bomb from a height of 40,000 feet. If the plane has a top speed of 600 mph, how far from the center of impact can the plane be when the bomb hits the ground? Can the plane's crew survive the shock of the bomb's explosion? What information do you need to answer these questions?

22. Suppose the gravitational attraction of the earth on an object varies inversely with the square of the distance between the object and the center of the earth.

(a) Show that an object M miles from the surface of the earth experiences an acceleration of $a = 32(4000)^2/(M - 4000)^2$ ft/sec/sec.

(b) Develop and analyze the model for free fall if acceleration is given as in part (a).

(c) According to the model of (b), how long should it take the moon to fall to the earth?

(d) Is the result of (c) relevant to an astronaut's trip home from a lunar exploration?

23. Using the inverse square law of Exercise 22, analyze the motion of a rocket fired from the earth's surface with a vertical velocity of v_0 ft/sec. Compute the maximum height that the rocket can reach. How large must v_0 be so that this height is greater than the distance from the earth to the moon? (Neglect the gravitational attraction of the moon on the rocket.)

24. Can a mathematical model of free fall be developed using position (y) and velocity (v) as the basic variables? Is it reasonable to adopt such an approach?

25. The main force producing the acceleration of an object in a vacuum is the force of gravity that causes the object to fall toward the earth. Archimedes (287−212 B.C.) discovered a force in the opposite direction: an object immersed in a medium (such as air or water or a gas) is buoyed up by a force equal to the weight of the medium displaced by the object.

(a) If m is the mass of the object and M is the mass of the medium displaced, show that the net force of gravity on the object is $32(m - M)$.

(b) Use the result of (a) to explain why a stone does not float in a lake of water, but a canoe does.

(c) Why does a balloon filled with helium rise in the air?

(d) Can buoyancy be ignored for a stone falling through the air? A snowflake?

26. From experimental observations, scientists have determined that air resistance of an object varies with the velocity of the object. Suppose that air resistance varies in direct proportion to velocity and is directed in the direction opposite to that of the velocity vector. Develop a mathematical model for the motion of an object hurled downward toward the earth from a height of y_0 feet through the atmosphere with an initial velocity of v_0 ft/sec. Take into account the forces of gravity and air resistance. Show that the motion may be modeled by the differential equation $dv/dt = 32 - kv$ where k is some constant.

27. Without solving the differential equation of Exercise 26, show that the velocity of a falling object subject to air resistance will "eventually" reach $32/k$ ft/sec regardless of the initial velocity v_0.

28. Develop a model for a freely falling object if the air resistance is proportional to the square of the velocity.

III. Discrete Models: Credit Cards and Populations

29. Suppose that your bank charges no interest.

(a) How large a monthly payment is needed to pay off a loan of \$1,000 in 2 years?

(b) How big a loan can you pay back in 4 years with monthly payments of \$31.25?

30. Provide a proof using mathematical induction that with 0% interest and a monthly payment of \$$p$, then the balance B_n of the loan after n months will be $B_n = B_0 - np$, where B_0 is the amount originally borrowed.

31. Provide a careful proof using the principle of mathematical induction that if $B_n = (1 + r)\,B_{n-1}$ for $n \geq 1$, then $B_n = (1 + r)^n B_0$ for all $n \geq 0$.

32. Suppose the *annual* interest rate (given as a decimal) is r, and a new balance is computed at n equally spaced times during a year.

(a) If the amount of the loan is B_0 and no payments are made, show that the balance of the loan after 1 year is given by $\left(1 + \frac{r}{n}\right)^n B_0$.

(b) What is the balance after t years?

(c) What happens to these balances in the limit as $n \to \infty$?

33. A bank charges 18% annual interest on credit card balances. What is the value of an original balance of \$1,000 after 1 year if the bank computes new balances once a year? Twice a year? Every 3 months? Every week? Every day?

34. Derive Eq. (27) from Eq. (26).

35. Derive Eq. (28) from Eq. (26).

36. To encourage new housing starts, the Federal Reserve lowers interests rates so that you can obtain a mortgage loan at a rate of 6% per year. What monthly payment is required if you want to pay off a loan of \$200,000 in 25 years? 20 years? 15 years?

37. With automobile loans at a 9% annual interest rate, you are contemplating borrowing \$25,000 to buy a new car. You can afford a monthly payment of \$300. How long will it take you to pay back the loan? How much in total interest will you have paid?

38. In our model of paying off your credit card, we assumed that the bank computed interest based on the previous month's charges and then subtracted your payment. Suppose the bank is required to deduct your payment from the balance before computing the interest. Show that for this model, the analogue of Eq. (21) would be

$$\begin{aligned} B_{new} &= B_{old} + r(B_{old} - p) - p \\ &= (1 + r)(B_{old} - p) = s\,(B_{old} - p) \end{aligned}$$

Compare the balances for the first 15 months under this model with the balances under the original model. Develop an analogue of Eq. (25).

39. Determine, if possible, analogues of Eq. (27) and Eq. (28) under the model presented in Exercise 36.

IV. Classification of Models

40. A mathematical model's predictions need not be forecasts of *future* events. They may be about phenomena that have occurred but of which observations have either not yet been made or for which such observations are unknown to the modeler. Find examples of such models. Consider instances in which it may not be possible to construct experiments—for example, a mathematical model for the frequency and intensity of political revolutions.

41. What kind of mathematical model is the usual set of axioms of plane geometry? What is being modeled?

V. Uses and Limitations of Mathematical Models

42. In what ways would you guess that high-speed electronic computers have affected the formulation of mathematical models of complicated phenomena? What benefits does the ability to do thousands of numerical calculations quickly confer? Are there any drawbacks?

SUGGESTED PROJECTS

1. Objects falling vertically in a resisting medium such as air, other gasses, or a liquid have a velocity v, which is often modeled by the differential equation $\frac{dv}{dt} = 32 - kv^{\alpha}$, where k and α are nonnegative constants. Investigate how the velocity, and, in turn, the distance above the ground, depends on the value of α when k is positive. In particular, does the object approach a "terminal velocity" well before hitting the ground? Can you obtain an

exact formula for v as a function of t for all values of α? Do "small" changes in α always produce "small" changes in the time of descent? Discuss how you might devise an experiment to determine the value of α.

2. We can often study the numerical behavior of a discrete dynamical system with a spreadsheet (e.g., *Excel* or *iWorks Numbers*) or a computer algebra system (e.g., *Maple* or *Mathematica*) using the basic iterative model. Implement the credit card balance model on one of these applications using the iterative equation

$$B(n) = (1 + r)\,B(n - 1) - p, \text{with } B(0) = 1000$$

Verify that the numbers in Table 1.3 are correct. Investigate what happens to the balance if you try different monthly payment schemes, such as (a) make a fixed payment \$$p$ payment only every other month or (b) start with a monthly payment of \$10 and increase the payment by a dollar every subsequent month.

3. If you rewrite the iterative equation of the population model as the difference equation $P_n - P_{n-1} = rP_{n-1} + p$, then the left-hand side represents the *change* in the population from one period to the next. For variables that change every instance, we often represent the change in value as a *derivative*. Treating the population in this fashion, show that a continuous analog to the difference equation is the differential equation $P'(t) = rP(t) + p$. Use calculus to solve this differential equation. (Hint: make the change of variable $y(t) = P(t) + \frac{p}{r}$.) Compare the predictions of the continuous model with those of the discrete model. Describe some real-world population growth situations that you believe would be more accurately described with a continuous model than a discrete one. When would a discrete model be more likely to provide significantly better, more realistic estimates?

4. Suppose you decide to build your savings around an investment in a certain stock. You make an initial investment of \$$B_0$ and an additional contribution of \$$p$ per month, but instead of a fixed rate r of interest, the stock may gain or lose value in each time period. If you assume that the stock's value could change, in some random fashion, positively or negatively by as much as 3% per month, then discuss why the model $B(n) = (1 + r)\,B(n - 1) + p$, where r is a random number between -0.03 and $+0.03$, is appropriate. Spreadsheets and computer algebra systems all have options to generate such random numbers. Implement this model in one of these software applications. Track your savings over a 10-year period. Redo the computations for many different time periods of 10 years. What is the average value for your savings after a decade? What was the largest amount you had after 10 years? The smallest? How much "variation" occurred?

You can find a listing of references and suggestions for additional reading on the book's website, www.wiley.com/college/olinick

Stable and Unstable Arms Races

CHAPTER 2

There is scarce truth enough alive to make societies secure, but security enough to make fellowships accurs'd. Much upon this riddle runs the wisdom of the world. This news is old enough, yet it is every day's news.

—William Shakespeare, *Measure for Measure*

I. The Real-World Setting

"Today, our troops have newer and better equipment, and their morale is high. The better armed they are, the less likely it is they will have to use that equipment. But if, heaven forbid, they are ever called upon to defend this nation, nothing would be more immoral than asking them to do so with weapons inferior to those of any possible opponent. . . .

None of the four wars of my lifetime came about because we were too strong. It is weakness that invites adventurous adversaries to make mistaken judgments.

[Ours] is the most peaceful, least warlike nation in modern history, We are not the cause of all the ills of this world. We are a patient and generous people. But for the sake of our freedom and that of others we cannot permit our reserve to be confused with a lack of resolve."

—Leader A

"Now let us turn to international affairs. One of the most important and insistent instructions of the . . . voters was, is and will remain, the instruction to safeguard peace like the apple of our eye and to ensure the security of our homeland. I can tell you that [we] have been strictly following this instruction, doing so in difficult circumstances.

You know that the past few years have seen a dramatic intensification of the policy of the more aggressive forces of . . . imperialism, a policy of blatant militarism, claim to world dominance, resistance to progress and violations of the rights and freedom of the peoples. . . .

All this compels us to attach the most serious attention to strengthening the country's defenses. [Our] people don't want an arms buildup, but rather the reduction of armaments by both sides. But we must take care to ensure sufficient security for our country, its friends and allies. This is precisely what is being done. And let everyone know that none of those given to armed ventures will catch us unawares, and no potential aggressor can hope to avoid devastating retaliation."

—Leader B

These actual quotations, from two recent world leaders (a U.S. President and a General Secretary of the former Soviet Union) of very different political persuasions exemplify the attitudes underlying the models we will study in this chapter. Although Ronald Reagan and Konstantin Chernenko disagreed on most issues, they both claimed the same rationale for maintaining a strong military with stockpiles of heavy armaments.

Imagine a nation, called Blue, whose people believe themselves desirous of a peaceful world. The leaders of this country share the people's fervent wish to achieve peace and avoid war. This is, of course, the public position of the vast majority of the world's governments; we assume that it is a sincerely held one.

The president of Blue and the other leaders of the government are not pacifists, however. They will not go out of their way to launch aggression, but they will not sit idly by if their country is attacked. The citizens of Blue share this attitude. They believe in self-defense and will fight to protect their nation and way of life. For this reason, they must be prepared to fight if necessity demands it.

The people of Blue feel that the maintenance of a large army and the stockpiling and improvement of weapons systems are purely defensive gestures. They have peaceful intentions and believe that if every nation were similarly solely concerned with self-defense, there would be no occasion for war. Aggressive acts are the cause of war; self-defense is not an aggressive act.

There is another large nation in this world, called Red. The people and leaders of Red share these same ideals, intentions, and ethical beliefs. They do not have hostile designs against anyone, but they are willing to fight to protect their homeland. The actions of the government of Blue to build and maintain armaments do not go unnoticed by the people of Red. Although the leaders of Blue continually proclaim peaceful intent, the weapons they have could be used to attack and destroy Red. The people of Red would consider their government derelict in its duty if it did not build up its armed forces for a secure defense. And so, the leaders of Red act accordingly.

Blue notes the ensuing increase in Red's arms expenditures. We know enough about the sensibilities of the people of Blue to realize that these increases will be seen as threatening to Blue's security and they will cry out for strengthening Blue's defensive forces.

In these past few paragraphs, we have sketched a highly simplified outline of an all-too-familiar international political problem. We wish to analyze the consequences of this situation to see what would happen if nations did behave in is manner.

The basic assumption is that of "mutual fear." The more that one nation arms, the more the other nation is spurred to arm. The more arms Blue accumulates, the more incentive is provided for Red to build up arms, and the more arms Red has, the more Blue is stimulated to arm. If the incentives to arm derive entirely from mutual fear and if neither side had any arms to start with, then an arms race would not start. But the slightest move on the part of one nation to build an army would initiate the whole vicious spiraling arms race.

This "mutual fear" model would predict that the armaments of both nations would continue to increase indefinitely as time went on. This cannot actually happen in the real world. No country has infinite resources. There is a limit to the amount of arms any nation can accumulate. When armament expenditures begin to absorb too large a portion of a nation's budget, there are protests within the country against raising the burden of arms costs still higher. Perhaps these limits force a leveling off of the arms race to some point of stability. Perhaps they do not.

The verbal analysis of this "mutual fear" situation is the sort we regularly see in newspaper and magazine articles and hear in the public speeches of our politicians. The verbal analysis cannot be carried much further. If we formulate a mathematical model of the situation, however, we can carry the analysis a good way and develop some consequences that may help us construct a more sophisticated theory of international politics.

In this chapter we will first present a very simple deterministic model of a "mutual fear" arms race. Examination of the consequences of this model will lead to the development of a more complex model that reflects more of the real-world situation. Mathematical analysis of this second model will yield new predictions about the course of arms races. We will then try to compare these predictions to an actual arms race.

II. Constructing a Deterministic Model

There are five major steps in the construction of a deterministic model of a situation that evolves over time:

1. Isolate and define the critical variables. In the models of arms races, the variables studied will be *armament expenditures*, their *rates of change*, and *time*.
2. Make assumptions about relationships among the critical variables and formulate them as equations or inequalities. In a "mutual fear" situation, we are supposing that the rate of change of arms expenditures with respect to time of one country will depend on the armament expenditures of the other country.
3. Apply mathematical analysis to the equations and inequalities. We hope to solve them or at least to discover information about the nature of the solutions. The analysis hopefully will give us new relations among the variables that are consequences of the assumed relations, but that were not immediately evident.
4. Interpret the results of the mathematical analysis as statements about the real world, and compare these with what actually happens in the real world.
5. Accept, discard, or improve the model. If the predicted relations are found to coincide closely with what we observe in the real-world situation, we may accept the model as a correct formulation. If the observations are in strong disagreement with the results of the mathematical analysis, discard the model and try a new approach. If some of the observations confirm the model and others do not, modify the assumptions about the relationships among the variables and formulate more accurate equations.

We will carry out these steps for a simple arms race model in the next section.

III. A Simple Model for an Arms Race

A. The Assumptions

Let x and y represent the yearly levels of armament expenditures of the two nations in some standardized monetary unit. These numbers are nonnegative and change with time.

Let t stand for time in years; assume that $t \geq 0$ and that the observation of the arms race begins at $t = 0$. The rates of change of x and y with respect to time are the derivatives $dx/dt = x'(t)$ and $dy/dt = y'(t)$.

Other important variables may occur to the reader. To keep this model simple, we will ignore, at least temporarily, other relevant quantities.

We wish to develop a simple model that reflects in some fashion the assumption of mutual fear: the more one nation arms, the more the other is spurred to arm. There are a variety of ways that this assumption could be formulated mathematically. It could be translated to mean that each country adjusts the level of its armaments to the level of the other country's. A more general approach would be that each country adjusts the rate of increase or decrease of its armaments in response to the level of the other's. To obtain a simple model, we will interpret this assumption to mean that each nation changes its expenditures at a rate directly proportional to the existing expenditure of the other nation.

Mathematically, the equations that state this assumption are

$$\frac{dx}{dt} = ay \tag{1}$$

$$\frac{dy}{dt} = bx \tag{2}$$

where a and b are positive constants. We do not claim, at this point, to know what numerical values a and b have. We do not need this knowledge to continue the mathematical analysis. The final results will show how the conclusions depend on the values of these parameters.

B. Mathematical Analysis

Eqs. (1) and (2) form what is called a *system* of first-order differential equations. A *solution* of the system is a pair of differentiable functions, $x = f(t)$ and $y = g(t)$, so that

$$f'(t) = a\,g(t) \text{ and } g'(t) = b\,f(t) \text{ for all } t \geq 0 \tag{3}$$

If the armament expenditures at time $t = 0$ of the two nations are x_0 and y_0, respectively, then we also insist that our solution satisfy these *initial conditions*—that is, f and g satisfy Eq. (3) and

$$f(0) = x_0,\ g(0) = y_0 \tag{4}$$

It can be shown that there is a unique pair of functions satisfying the conditions of Eqs. (3) and (4). The proof is outlined in the Exercises. We want to derive some information about the nature of the solution functions f and g.

In the first place, the fact that the parameters a and b are positive implies that the derivatives dx/dt and dy/dt are nonnegative. Thus $f'(t) \geq 0$ and $g'(t) \geq 0$. From elementary calculus, we may conclude that f and g are nondecreasing functions of t.

Second, differentiate each side of Eq. (1) with respect to t and obtain

$$\frac{d^2x}{dt^2} = a\frac{dy}{dt} = \text{ that is, } f''(t) = a\,g'(t) \tag{5}$$

so that, the second derivative of f is also nonnegative. The geometric conclusion of this observation is that the graph of f is concave up. By differentiating Eq. (2), we may conclude by a similar argument that the graph of g is also concave up.

We can obtain more information about the relationship between the solution functions by making use of the chain rule to write

$$\frac{dy}{dx} = \frac{dy/dt}{dx/dt} = \frac{b}{a}\,\frac{x}{y} \tag{6}$$

Rewriting Eq. (6) in differential form and integrating, we obtain

$$\int y\,dy = \int \frac{b}{a}x\,dx \tag{7}$$

or $\frac{y^2}{2} = \left(\frac{b}{a}\right)\frac{x^2}{2} + K$, which gives

$$y^2 = \left(\frac{b}{a}\right)x^2 + C \tag{8}$$

where $C = 2K$ is the constant of integration. The value of C is obtained by substituting $t = 0$ into Eq. (8) and using the fact that $y(0) = g(0) = y_0$, while $x(0) = f(0) = x_0$. What is more important for our understanding of this simple arms race model is that Eq. (8) is the equation of a hyperbola in the plane with straight-line asymptote $y = \sqrt{b/a}\,x$. The graph of the hyperbola in the plane is sketched in Fig. 2.1.

Suppose that at some time the x and y values determine a point in the interior of the first quadrant on the upper branch of the hyperbola. Then, as time continues, the points $(x(t),y(t))$ remain on this branch, moving in a northeasterly direction, since (by Eq. (6)) the

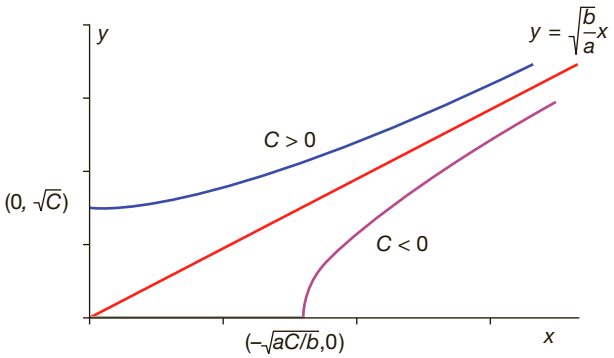


FIGURE 2.1 $C = y_0^2 - (b/a)x_0^2$ The two curves illustrate the possibilities for the simple arms race model.

slope of the tangent line is positive. This branch lies above the asymptote line, so that $y > \sqrt{b/a}\,x$. Eq. (1) then gives

$$\frac{dx}{dt} = ay > a\sqrt{b/a}\,x = \sqrt{ab}\,x$$

or

$$\frac{dx}{dt} > \sqrt{ab}\,x$$

Thus, as the x-coordinate increases, so does the velocity of the horizontal motion. The motion along the hyperbola is speeding up and the values of x and y will increase without bound as time increases—that is,

$$\lim_{t\to\infty} x(t) = \lim_{t\to\infty} y(t) = \infty$$

The same result can be established for motion along the lower branch of the hyperbola.

C. The Conclusions

It is possible, for those who know a little more about differential equations, to find explicit formulas for the solution f and g as functions of t (see Exercise 11). Even without doing this, we have enough information to analyze the qualitative consequences of this simple arms race model: both nations will spend more and more money on armaments as time proceeds, with no limit on the expenditures.

Note that this mathematical prediction is consistent with part of the verbal analysis of this "mutual fear" model. The mathematical prediction of indefinitely large expenditures, however, violates commonsense observations that there must be a finite limit to the expenditures. We should modify the model to reflect this observation. We explore such an improved model in the next section.

IV. The Richardson Model

A. The Assumptions

For the refined model, we begin with the premise of mutual fear expressed by the assumption that the rate of change of armament expenditures of each country is directly proportional to the expenditures of the other country. We will also attempt to include the "limiting factors" discussed above. We can do so by assuming that excessive armament expenditures present a drag on the nation's economy so that the actual level of expenditures depresses the rate of expenditure changes. A mathematical way of expressing this is to assume that the rate of change for a nation is directly and negatively proportional to its own expenditures. More precisely, a simple refinement of the original model would be the system of differential equations

$$\frac{dx}{dt} = x'(t) = ay - mx \tag{9}$$

$$a, b, m, n > 0$$

$$\frac{dy}{dt} = y'(t) = bx - ny \tag{10}$$

Where a, b, m, and n are all positive constants. Rather than proceed to a mathematical analysis of this model, we will consider a model that is a further refinement of this one. The idea of the second refinement is that some people argue that the cause of increasing arms expenditures is not mutual stimulation but permanent underlying grievances of each country against the other. To satisfy these analysts, Lewis F. Richardson (1881–1953) proposed the following arms race model:

$$\frac{dx}{dt} = x'(t) = ay - mx + r \tag{11}$$

$$\frac{dy}{dt} = y'(t) = bx - ny + s \tag{12}$$

where a, b, m, and n are positive constants and r and s are constants that may have any sign.

Assignment of a positive value to r or s indicates that there is a grievance by one country against the other that spurs it to accumulate arms. If we assign a negative value to one of these parameters, however, then we are asserting that there are underlying feelings of goodwill that tend to diminish perceptions of threat and hence to decrease dependence on arms.

The differential equations of the Richardson model then assert that the rates of arms expenditure increase of one nation depend positively on the expenditure level of the other country, negatively on the country's own expenditures, and positively on underlying grievances. The values assigned to the six parameters measure the extent of these effects.

The Richardson model is quite a flexible one. Analysts who differ in their beliefs as to the relative importance of the three determinants of rate of change may choose values for the parameters to reflect their preferences. If one of the three factors is believed irrelevant to changes in expenditures, then the corresponding parameter can simply be set equal to zero. If, for example, all the constants except r and s are zero, then only the grievances are considered as contributing to changes in armaments.

Thus, you can use the Richardson model if you believe that an arms race is a self-stimulating process or if you believe that self-stimulation has nothing to do with accumulation of arms. Assigning nonzero values to all the constants gives a model that contains the essential assumptions of our verbal description:

1. Arms accumulate because of mutual fear.
2. There is resistance within society to ever-increasing arms expenditures.
3. There are considerations independent of expenditure levels that contribute to the buildup of armaments.

We shall see that the nature of solutions to the system of differential equations of the Richardson model depends not on the precise values of the parameters, but rather on their relative magnitudes and the signs of the "grievance" terms, r and s.

B. Elementary Analysis of the Model

Let us suppose that the arms race begins at time $t = 0$ and that the differential equations (11) and (12) are valid for all time $t \geq 0$. Suppose also that at $t = 0$, Blue and Red are respectively spending at annual rates of x_0 and y_0 monetary units on armaments. It can be shown that there is a unique pair of differentiable functions of t, $x = f(t)$, and $y = g(t)$ such that

$$f'(t) = a\,g(t) - mf(t) + r \tag{13}$$

$$g'(t) = bf(t) - ng(t) + s \tag{14}$$

$$f(0) = x_0,\ g(0) = y_0 \tag{15}$$

It is possible to solve the system of differential equations to obtain f and g explicitly as functions of t, the six parameters, and the initial expenditures. We outline the necessary mathematical procedures in Section VI below. In this section, we will analyze the system with the tools you learned in elementary calculus.

The techniques and concepts of calculus shed light on an important aspect of arms races: *stability*. Some of the terms in the Richardson equations act to increase expenditures, while others put a brake on spiraling costs. Is it possible that the combined effect of all the terms will force arms expenditures ultimately to become "stabilized"? Will the level of expenses approach or remain at some fixed, constant amount?

The mathematical requirement that a nation's expenditures stay constant is that the rate of change be zero. We say that the arms race *stabilizes* when both nations reach a level of constant expenditures. For stabilization to occur, then, both rates of change must become zero—that is,

$$\frac{dx}{dt} = 0 = \frac{dy}{dt} \tag{16}$$

From the Richardson equations, this is the same as demanding that

$$ay - mx + r = 0 \tag{17}$$

and

$$bx - ny + s = 0 \tag{18}$$

In our analysis, we will assume that the parameters are all nonzero. The exploration of the special cases in which some of the parameters are assigned values of zero will be left for the Exercises.

Accordingly, we may rewrite Eq. (17) as

$$y = \frac{m}{a}x - \frac{r}{a} \tag{17a}$$

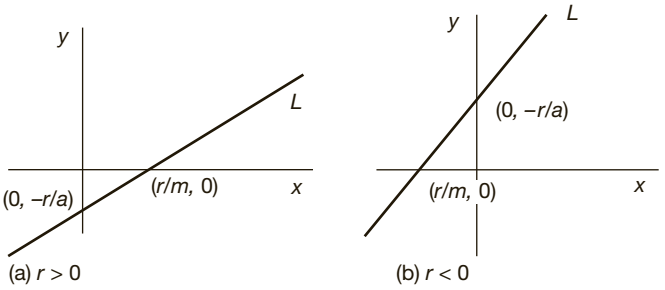


FIGURE 2.2 The line L: $y = (m/a)x - (r/a)$. At points along this line, $dx/dt = 0$.

This equation represents the straight line L with slope m/a, y-intercept $(0, -r/a)$ and x-intercept $(r/m, 0)$; see Fig. 2.2.

It is useful to think of the Richardson system of differential equations as the equations of motion of a particle in the (x, y)-plane. The first equation gives the horizontal component of velocity and the second equation gives the vertical component of velocity. The equations assert that the velocity components are functions purely of the x- and y-coordinates of a point and of certain constants. All we really require here is to recall simple facts, such as that if dx/dt at some point is positive, then $x(t)$ is increasing at this point, so that the particle will tend to move to the right; if dx/dt is negative, the particle tends to the left. If dy/dt is positive (negative), then the particle will move up (down).

If at some instant of time, the levels of expenditures (x_1, y_1) of Blue and Red happen to coincide with a point on L—that is, $ay_1 - mx_1 + r = 0$—then dx/dt at (x_1, y_1) will be zero. The expenditures of Blue at that moment will not be changing. Of course, dy/dt at this point is likely to be nonzero, so the level of expenditures may move up or down at that instant toward a point not on L.

The line L is called the *optimal line* for Blue. We shall see that the Richardson model implies that Blue is continuously changing its expenditures levels to bring them closer to the optimal line.

We wish to explore the limiting behavior of the Richardson arms race model as t gets large. We can distinguish three cases:

1. A runaway arms race: $x \to \infty$ and $y \to \infty$
2. Mutual disarmament: $x \to 0$ and $y \to 0$
3. A stable arms race: $x \to x^*$ and $x \to y^*$ for some positive numbers x^* and y^*

If there is a stable arms race, it is easy to determine the values of x^* and y^*. We consider the lines L and L' where $dx/dt = 0$ and $dy/dt = 0$, respectively. The line L' is Red's optimal line. The two lines intersect in a point (x^*, y^*); see Fig. 2.3. At this level of armament expenditures, the rates of increase for both nations are zero and will remain at this level, (x^*, y^*), which will be called the *point of stability*.

The optimal line L divides the plane into two open half-planes; see Fig. 2.4. One half-plane consists of all the points to the "right" of L, and the other is made up of all the points to the "left" of L. More formally, a point P *lies to the right* of a line L if the horizontal line through P hits L at a point with smaller x-coordinate while P lies *above* L if the vertical line through P hits L at a point with smaller y-coordinate.

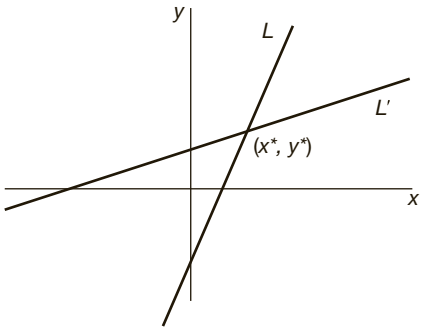


FIGURE 2.3 The intersection of optimal lines at (x^*, y^*) the point of stability. At this point, both derivatives dx/dt and dy/dt are zero.

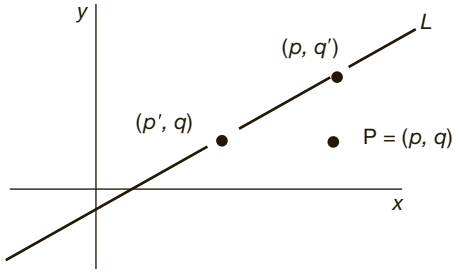


FIGURE 2.4 The point (p, q) lies below and to the right of the line L.

Suppose that (x_1, y_1) is any point not on the optimal line L, and let (x_2, y_1) be the corresponding point on L. Then $ay_1 - mx_2 + r = 0$, while the derivative dx/dt at (x_1, y_1) has value

$$\begin{aligned} x'(x_1, y_1) &= ay_1 - mx_1 + r \\ &= ay_1 - mx_1 + r - 0 \\ &= ay_1 - mx_1 + r - (ay_1 - mx_2 + r) \\ &= m(x_2 - x_1) \end{aligned} \tag{19}$$

Now we see that dx/dt at (x_1, y_1) is positive exactly when $x_2 > x_1$, and this occurs exactly when (x_1, y_1) lies to the left of L. The derivative is negative, similarly, exactly when (x_1, y_1) lies to the right of L. Thus, if (x_1, y_1) lies to the left of L, then the horizontal motion at that moment is toward L, whereas if (x_1, y_1) lies to the right of L, the horizontal motion is also toward L. In either case, the Richardson model implies that Blue is always adjusting its expenditures to move them toward the optimal line—that is, Blue is trying to stabilize its arms expenses. See Fig. 2.5.

Similar analysis yields corresponding results for Red's optimal line L' along which $dy/dt = 0$. The derivative is negative at any point above L' and is positive at any point below L'. Thus, the vertical motion is always toward L'; the nation of Red always adjusts its expenditures toward its optimal line.

C. Does Stability Occur?

We have just seen that the Richardson model predicts horizontal movement toward L and vertical movement toward L'. In other words, if the initial level of armament expenditures is

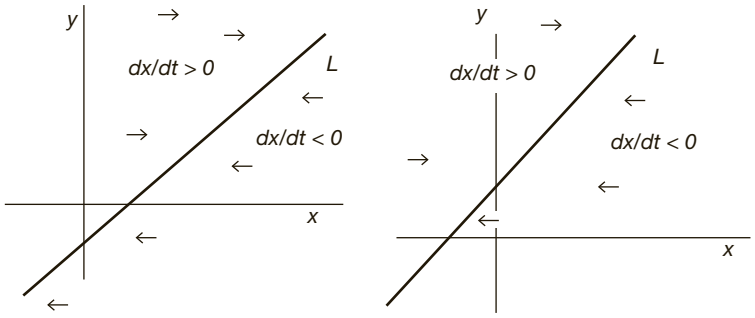


FIGURE 2.5 Blue adjusts expenditures toward the line L. On L, $dx/dt = 0$.

(x_0, y_0), then Blue will change its expenditures to bring them closer to its optimal line and so will Red. In this section, we will investigate what effects these motions have on the possibility of stabilizing the arms race.

If x and y represent armament expenditures, then we attach no meaning, for the present, to these variables' being negative. We consider, then, only the behavior of solutions of the Richardson model when x and y are nonnegative. We examine only the portion of the graph relating y and x that lies in the first quadrant.

D. Mutual Grievances

Let us first investigate the arms race in which each side has a permanent underlying grievance against the other side. Mathematically, this means we will assume that the parameters r and s are both positive.

In this situation suppose that Blue and Red are completely disarmed, so that the initial expenditure level is (0, 0). According to the Richardson model, the rates of change of expenditures at this instant would be

$$\frac{dx}{dt} = a0 - m0 + r = r > 0 \tag{20}$$

and

$$\frac{dy}{dt} = b0 - n0 + s = s > 0 \tag{21}$$

so that each nation would start arming itself.

Can this system be stable? In the case in which r and s are both positive, the point of stability (x^*, y^*) will lie in either the first quadrant or the third quadrant (see Exercise 18). We will start with the case that (x^*, y^*) lies in the first quadrant. Then the lines L and L' will be as pictured in Fig. 2.6. These lines split the first quadrant into four regions. Label them, counterclockwise, I, II, III, IV, so that the origin is in region III.

If the initial armament expenditures are at (0, 0), then, as we have seen, both x and y will increase. The net result will be to move the expenditures toward a point (x_1, y_1) deeper

Region	$\frac{dx}{dt}$	$\frac{dy}{dt}$
I	< 0	< 0
II	> 0	< 0
III	> 0	> 0
IV	< 0	> 0

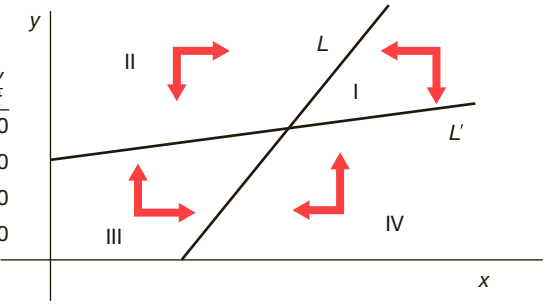


FIGURE 2.6 The optimal lines split up the first quadrant into four regions. The arrows indicate the horizontal and vertical directions of motion in each region.

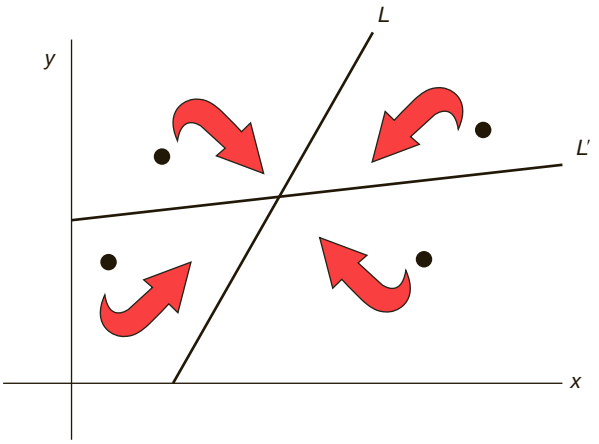


FIGURE 2.7 The stable case. Possible initial levels (x_0, y_0) are indicated with solid circles.

in region III, closer to the stability point (x^*, y^*) a short time later. If (x_1, y_1) is any point in region III, then, again, dx/dt and dy/dt will be positive and the motion will still be in a "northeasterly" direction toward (x^*, y^*). If at some instant, the motion carries the particles to the piece of the line L separating regions III and IV, then at such a point $dx/dt = 0$ while dy/dt is positive; the resulting motion is vertical and returns to region III. Similarly, at every point on the part of L' separating regions II and III, $dy/dt = 0$ while dx/dt is positive; again motion is back into region III. No matter where in region III the initial level of expenditures (x_0, y_0) is, the long-term behavior of the arms race is movement toward (x^*, y^*) and stability results. See Fig. 2.7.

The initial expenditure levels (x_0, y_0) could, of course, be at a point in one of the other three regions of the first quadrant. It is easy to check whether in any of these cases, the movement is again toward the stable point (x^*, y^*). In Fig. 2.7, this is shown for several different initial levels. Thus, whenever the optimal lines intersect in the first quadrant and the grievance terms r and s are positive, we have a stabilizing arms race. Any deviation from the "Balance of Power" point (x^*, y^*) will tend to be corrected.

E. Stable Point in Third Quadrant

To continue the analysis for the case when the parameters r and s are positive, consider what happens if (x^*, y^*) is in the third quadrant. Then the relationship of the optimal lines L and L' is the one pictured in Fig. 2.8.

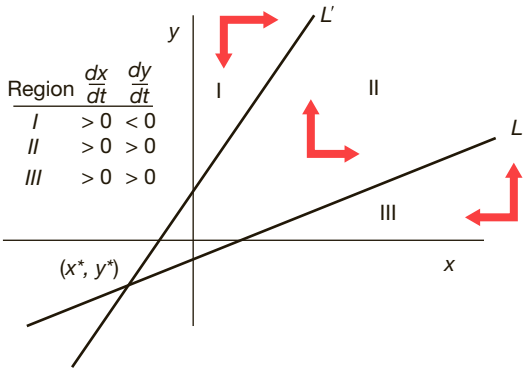


FIGURE 2.8 Runaway arms race. Grievance terms r and s are both positive.

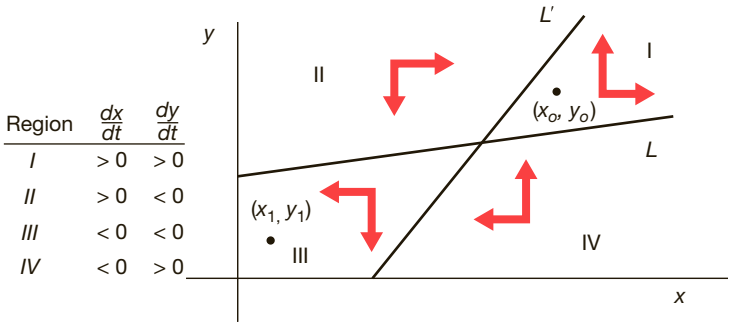


FIGURE 2.9 An ambiguous case. Both grievance terms r and s are negative.

Since the stable point is unobtainable through any positive levels of armament expenditures, a stabilizing arms race is not possible. The first quadrant is split into three regions as indicated in Fig. 2.8. Investigation of the signs of dx/dt and dy/dt in these regions shows that no matter the initial level of expenditures, the motion of the system eventually carries expenditures into the second region. Once the expenditures reach a point in this second region, both x and y values continue to increase and the expenditures reach indefinitely high levels; there is a runaway arms race no matter what the initial expenditures were.

In the Exercises, you will prove that the situation just discussed can only occur if $mn < ab$—that is, the "combined" effect of the braking terms is not enough to offset the terms that measure mutual stimulation to increase arms expenditures.

F. The Bad Effect of Good Will

In this section we will show that there are some cases in which the nature of the ultimate behavior of the Richardson model depends upon the initial level of expenditures. Look at the situation pictured in Fig. 2.9. This can occur only if at least one of the "grievance" terms, r or s, is negative—that is, only if at least one of the nations has feelings of "goodwill" toward the other.

Suppose the initial level of expenditures is at the point (x_0, y_0) in Fig. 2.9. Expenditures will not remain at this point, because it is not the point of stability. Since (x_0, y_0) is to the left of L and below L', the x-coordinate and y-coordinate will both increase and the expenditures will

move to a new point in region I farther from the stable point. The arms expenditures for both nations will increase indefinitely as time goes on. There is a runaway arms race.

On the other hand, suppose that the initial level is at (x_1, y_1) in region III of Fig. 2.9. Now we are to the right of L and above L', so both coordinates will decrease, and we will move to another point in region III farther from the stable point. We are headed toward mutual disarmament.

If the initial point is in region II or IV, then the analysis is more complicated. These cases will be discussed in some detail in later sections. The basic result, however, is easy to state. The ultimate behavior is either a runaway arms race or total disarmament, depending on whether we move first into region I or region III. This is determined by the location of the initial expenditure levels.

There is a rather ironic situation here. If underlying grievances exist (both r and s positive), then a stable arms race may result, independent of how high the initial level of expenditures is or how great the disparity in expenditures of the nations is at the start. When the underlying feelings are of goodwill (negative values for r and s), then a runaway arms race is an alternative to disarmament, and the eventual outcome very much depends on where you start.

This ambiguous case arises when the lines L and L' intersect in the first quadrant. If the point of intersection happens to lie in the third quadrant, and r and s are both negative, then the arms race becomes a march to mutual disarmament regardless of the location of the initial point (see Exercise 23).

Direction fields Computer software packages make it possible to visualize the dynamics of a system of differential equations $dx/dt = f(x, y)$, $dy/dt = g(x, y)$ in another way. We draw small pointed line segments with a slope $\dfrac{dy}{dx} = \dfrac{g(x,y)}{f(x,y)}$ at any desired point (x_i, y_i). Each such arrow is then tangent to the solution at the point (x_i, y_i). The set of all these line segments is called the *direction field* or *slope field*. By plotting the direction field of the system, you can get a good qualitative feel for the solution and its properties. It's tedious to draw very many of these line segments by hand, but a computer can carry this task quite quickly. Figs. 2.10 and 2.11 show the direction field for two Richardson arms race models.

Summary of results The eventual outcome of an arms race that follows the Richardson model depends upon the relative sizes of the parameters a, b, m, and n and the signs of r and s. We have noted the following typical cases:

Case 1: If $mn - ab$ is positive, and r and s are positive, then there will be a stabilized arms race.

Case 2: If $mn - ab$ is negative, and r and s are positive, there will be a runaway arms race.

Case 3: If $mn - ab$ is positive, and r and s are negative, then there will be total disarmament.

Case 4: If $mn - ab$ is negative, and r and s are negative, then the situation is ambiguous. There will either be disarmament or a runaway arms race, depending on the initial level of expenditures.

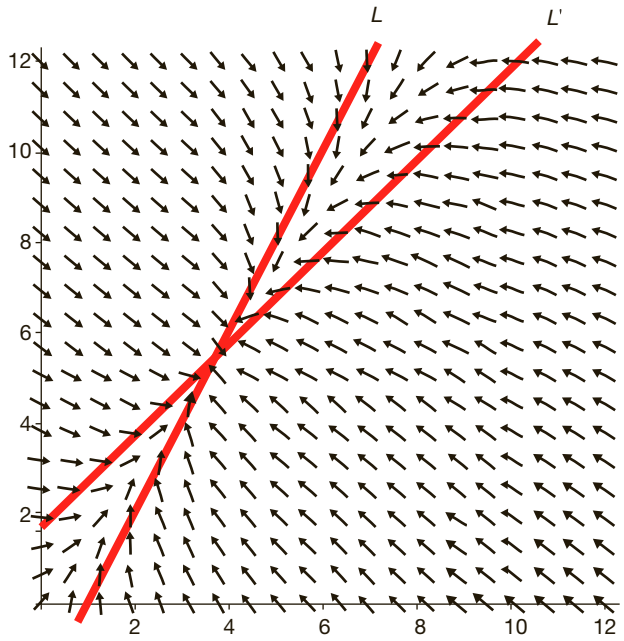


FIGURE 2.10 Direction field for the arms race model $dx/dt = 3y - 6x + 6$, $dy/dt = 4x - 4y + 7$.

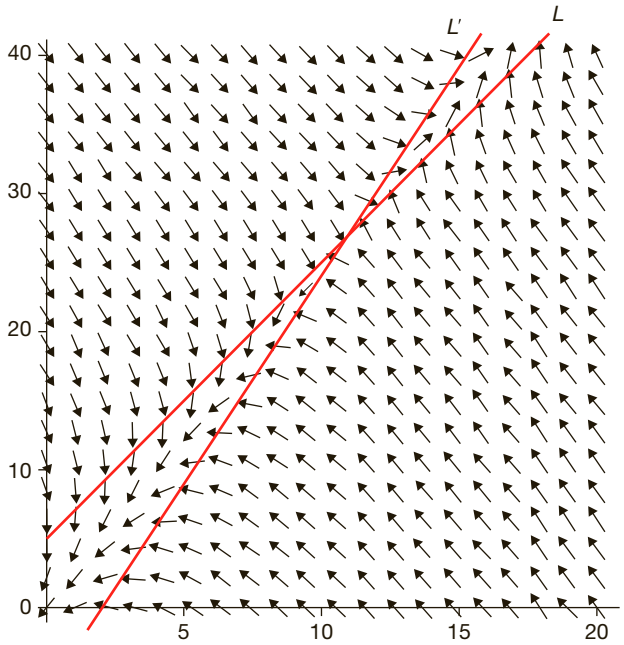


FIGURE 2.11 Direction field for the arms race model $dx/dt = 1y - 2\,x - 5$, $dy/dt = 6x - 2\,y - 12$.

G. Further Analysis of the Richardson Model

Suppose that you wish to study the Richardson model in a situation where investigation indicates that the values of the various parameters should be

$$a = 1, \quad m = 2, \quad r = -5, \quad b = 6, \quad n = 2, \quad s = -12$$

The optimal lines are L: $y - 2x - 5 = 0$ and L': $6x - 2y - 12 = 0$. The two lines intersect at the stable point $(x^*, y^*) = (11, 27)$.

Since the grievance terms r and s are negative and $mn - ab = 4 - 6$ is negative, you have the ambiguous case in which ultimate behavior of the system depends on the location of the initial level of expenditures.

Observation of this particular system at time $t = 0$ shows that the initial level is $(x_0, y_0) = (15, 15)$. At this point, the derivatives are given by

$$x'(x_0, y_0) = x'(15, 15) = 15 - 2(15) - 5 = -20$$

and

$$y'(x_0, y_0) = y'(15, 15) = 6(15) - 2(15) - 12 = 48$$

The signs of these derivatives indicate that the initial point is in region IV of Fig. 2.9. You cannot tell from any of the mathematical analysis yet presented what the ultimate behavior of this particular arms race will be. In this section, we will present two techniques—of general application in the solution of systems of differential equations—that will help determine the outcome.

The Euler method This method, introduced by the great Swiss mathematician Leonhard Euler (1707–1783), is based on a simple geometric interpretation of the derivative.

Suppose that $u = f(t)$ is a differentiable function of t and that the value of the function and its first derivative are known at a number t_0. We wish to approximate the value of the function at a nearby number $t_0 + \Delta t$. Direct computation may be quite difficult. Note, however, that the graph of the tangent line to the curve $u = f(t)$ stays close to the curve near a point of tangency $\big(t_0, f(t_0)\big) = (t_0, u_0)$. The slope of the tangent line is given as $f'(t_0) = u'(t_0, u_0)$. It is a simple matter to use the equation of the tangent line to find the point on that line with the first coordinate equal to $t_0 + \Delta t$. If Δt is small, then the second coordinate of this point is a good approximation to the value $f(t_0 + \Delta t)$ since the tangent line will not wander far from the curve (see Fig. 2.12). The smaller Δt is, of course, the better the approximation will be. Analytically, the actual change in the function from t_0 to $t_0 + \Delta t$ is

$$\Delta u = f(t_0 + \Delta t) - f(t_0)$$

The approximation is that Δu is roughly equal to $f'(t_0)\,\Delta t$ so that

$$f(t_0 + \Delta t) \sim u_0 + f'(t_0)\,\Delta t$$

which may also be written as

$$f(t_0 + \Delta t) \sim u_0 + u'(t_{0,}\, u_0)\,\Delta t$$

where $\sim$ is a symbol representing "approximately equal to."

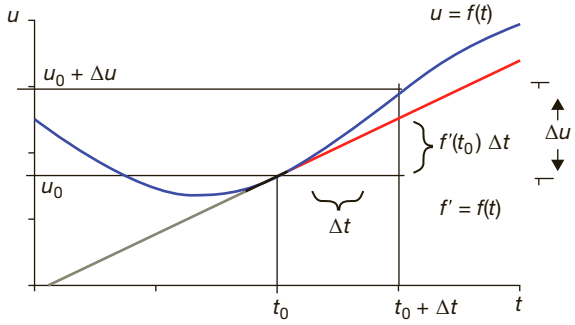


FIGURE 2.12 The change in height of the graph of a function near the point of tangency is approximated by the change in height of the tangent line.

Example

Suppose $u = f(t) = \sqrt{t}$ and $t_0 = 4$. Then $u_0 = \sqrt{4} = 2$ and $f'(t) = \frac{1}{2\sqrt{t}}$ so that $f'(t_0) = f'(4) = u'(4, 2) = \dfrac{1}{2\sqrt{4}} = \dfrac{1}{4}$. The approximation that is made here is

$$\sqrt{4 + \Delta t} \sim 2 + \frac{1}{4}\Delta t$$

If we wish to compute $\sqrt{4.41}$, for example, then we take $\Delta t = .41$. The approximate value is $2 + \dfrac{1}{4}(.41) = 2.1025$. The actual value of $\sqrt{4.41}$ is 2.100. The approximation here is quite good.

Many calculus texts contain detailed discussion of this method of "increments" for approximations. See, for example, Section 2.8, "Linear Approximations and Differentials," in Swokowski, Olinick, Pence, *Calculus*, 6th ed., Boston: PWS, 1994.

The method of increments is the basis for Euler's technique of approximating solutions to differential equations. In the context of the Richardson arms race model, suppose that the initial level of expenditures is (x_0, y_0). Then the rate of change for Blue is $x'(x_0, y_0) = ay_0 - mx_0 + r$ and for Red it is $y'(x_0, y_0) = bx_0 - ny_0 + s$. In a short time interval Δt, the amount Δx that the arms expenditures for Blue will change is approximately $x'(x_0, y_0)\Delta t$. The change of expenditures for Red during this same time interval is denoted Δy and is approximately equal to $y'(x_0, y_0)\Delta t$.

Thus at time $t_0 + \Delta t = 0 + \Delta t = \Delta t$, the new expenditure levels will be at the point $P_1 = (x_1, y_1)$ The coordinates of this point are estimated by the method of increments to be

$$x_1 \sim x_0 + x'(x_0, y_0)\,\Delta t$$

and

$$y_1 \sim y_0 + y'(x_0, y_0)\,\Delta t.$$

If we choose Δt to be small, then the estimated coordinates will be quite close to the actual coordinates at time Δt.

Once a new point P_1 has been estimated, it may be treated as the initial point of the system and the method of increments can be applied again. This may be done as often as you like. The general formula for estimating successive points would be

$$P_{i+1} = (x_{i+1}, y_{i+1}) \quad \text{where} \quad \begin{cases} x_{i+1} = x_i + x'(x_i, y_i)\Delta t \\ y_{i+1} = y_i + y'(x_i, y_i)\Delta t \end{cases} \tag{22}$$

This formula for P_{i+1}can be applied to any system of differential equations in which dx/dt and dy/dt are given as explicit functions of x and y. For the Richardson arms race model, the formula becomes

$$P_{i+1} = (x_{i+1}, y_{i+1}) \quad \text{where} \quad \begin{cases} x_{i+1} = x_i + (ay_i - mx_i + r)\Delta t \\ y_{i+1} = y_i + (bx_i - ny_i + s)\Delta t \end{cases} \tag{23}$$

For the particular Richardson model we have been discussing in this section, the formula reduces to

$$P_{i+1} = (x_i + (y_i - 2x_i - 5)\,\Delta t,\, y_i + (6x_i - 2y_i - 12)\Delta t) \tag{24}$$

$$P_0 = (15, 15)$$

If Δt is chosen to be .01, then repeated use of Eq. (24) yields the following data:

I	x_i	y_i	$x'(x_i, y_i)$	$y'(x_i, y_i)$
.00	15	15.	−20.	48.
.01	14.8	15.48	−19.12	45.84
.02	14.608800	15.938400	−18.279200	43.776000
.03	14.426008	16.376160	−17.475856	41.803728
.04	14.251249	16.794197	−16.708302	39.919102
.05	14.084166	17.193388	−15.974945	38.118222
. . .				
.10	13.352434	18.934690	−12.770177	30.245223
.20	12.298813	21.409885	−8.187741	18.973108
.30	11.621573	22.958293	−5.284853	11.812852
. . .				
.89	10.407465	25.169142	−0.645788	0.106505
.90	10.401007	25.170207	−0.631807	0.065628
.91	10.394689	25.170864	−0.618515	0.026407
.92	10.388504	25.171128	−.60588	−.001123

Since both derivatives at P_{92} are negative, the point P_{92} is in region III and, by our previous analysis, we may conjecture that the ultimate result of this particular arms race will be total disarmament.

This incremental method yields a definite and believable result. It is difficult to ascertain the ultimate behavior of an arms race in this way, however, unless we are willing to make a great many computations. In these computations, moreover, there is a slight error at each step, because the incremental method yields only approximations. This error may build up over the many steps necessary in the computation so that the actual position of the expenditure levels may be quite far away from the estimated one. The trick is to make Δt small enough that the accumulated error is small, but large enough that the number of calculations before reaching region I or III is reasonable.

Rather than pursue this method any further at this stage—or examine the many refinements mathematicians have developed for the numerical solution of systems of differential equations—we will describe another method of answering the stability question. This method will accurately predict the outcome of the arms race, will not involve extensive calculations, and is mathematically defensible.

The point-slope method The motivation behind this method derives from trying to answer the question "Is it possible that the movement of arms expenditures levels could be motion *along a straight line* toward the stable point?"

Let $P_0 = (x_0, y_0)$ be the initial point, and consider the straight line L^* through P_0 and the stable point $S = (x^*, y^*)$. The slope of this line is

$$\frac{y^* - y_0}{x^* - x_0}$$

If the motion of the point representing armament expenditures is given by the Richardson model, then the chain rule gives

$$\frac{dy}{dx} = \frac{dy/dt}{dx/dt} = \frac{bx - ny + s}{ay - mx + r} \tag{25}$$

This fraction measures the slope of the line tangent to the curve along which the point is moving at the instant when the point is at (x, y). See Fig. 2.13.

Suppose that this direction of motion happens to be along L^*, and furthermore assume that every point on L^* has this property—namely, the slope of the line L^* through S and the point is equal to the derivative dy/dx evaluated at the point. Then, if the initial point is on L^*,

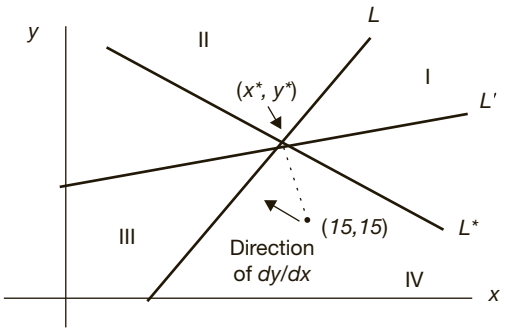


FIGURE 2.13 Illustration of the point-slope method.

all subsequent points (hence, all subsequent levels of arms expenditures) will be on L^*. This is true because at $t = 0$, the point is on L^* and is moving along L^*, and any instant later it has reached another point on L^*, where dy/dx determines the new direction of motion—and this direction is again along L^*.

We can find the equation of such a line L^* by using the property that we have assumed for it: a point (x, y) is on L^* if and only if dy/dx at that point is equal to the slope of the line from S to the point—that is,

$$\frac{bx - ny + s}{ay - mx + r} = \frac{y^* - y}{x^* - x} \tag{26}$$

This equation may be solved for y as a function of x. In the example under consideration, a point (x, y) is on L^* if and only if

$$\frac{6x - 2y - 12}{y - 2x - 5} = \frac{27 - y}{11 - x} \tag{27}$$

This equation is equivalent to

$$(6x - 2y - 12)(11 - x) = (27 - y)(y - 2x - 5)$$

Perform the indicated multiplication and rearrange terms to obtain

$$y^2 - 54y - (6x^2 - 132x - 3) = 0 \tag{28}$$

Eq. (28) may be treated as a quadratic equation in y with constant term $6x^2 - 132x - 3$. The two solutions are

$$y = 27 + \sqrt{6}(x - 11) \quad \text{and} \quad y = 27 - \sqrt{6}(x - 11) \tag{29}$$

The equations in Eq. (29) are equations of straight lines. Both lines pass through the stable point (11, 27). The first equation represents a line of positive slope that runs through regions I and III and is not of interest to us. The second line has negative slope and lies in regions II and IV. This is the equation of the desired line L^*.

Consider any point on $L*$ in region IV. Here dx/dt is negative and dy/dt is positive, so that the direction of motion is "northwesterly." Since the point is on the special line L^*, the motion is along L^*. Starting at any such point, the motion will be toward the stable point as time progresses. A similar consideration of the signs of dx/dt and dy/dt shows that if motion starts at a point of L^* in region II, subsequent motion is along L^* toward the stable point.

Now a point (x, y) will be on L^* exactly when $y + \sqrt{6}(x - 11) - 27 = 0$. This line divides the plane into two regions, one below the line and one above it, for which $y + \sqrt{6}(x - 11) - 27$ is negative and positive, respectively. These correspond to the inequalities

$$\frac{6x - 2y - 12}{y - 2x - 5} > \frac{27 - y}{11 - x} \tag{30}$$

for points below L^*, and

$$\frac{6x - 2y - 12}{y - 2x - 5} < \frac{27 - y}{11 - x} \tag{31}$$

for points above L^* in regions II and IV.

If (x, y) is a point in the plane above L^*, and we connect that point to the stable point with a straight line, then the direction of motion from this point will be above this line, and closer to region I. The ultimate motion will carry arms expenditure levels into region I, and the arms race will escalate without limit.

If, on the other hand, the initial point of the arms race is below L^* then the direction of motion away from the point is below the line connecting that point to the stable point. The motion is toward region III, which the system will eventually enter. Total disarmament will result.

For the particular example under consideration with initial point (15, 15), the direction of motion at this point is

$$\frac{6(15) - 2(15) - 12}{15 - 2(15) - 5} = \frac{48}{-20} = -2.4$$

while the slope of the line from (15, 15) to the stable point (11, 27) is

$$\frac{27 - 15}{11 - 15} = -3$$

Since -2.4 is greater than -3, the direction of motion is below the line from (15, 15) to (11, 27), so the movement is away from the stable point and toward region III. The result is eventual disarmament.

If we consider a particular case of the Richardson arms race model, with assigned values for the parameters and the initial expenditures, then the set of all later expenditures traces out a curve in the plane. This curve may move toward the stable point, or toward the origin, or it may simply assume larger and larger values of both coordinates without limit as time goes on. Equipped with the point-slope method and the analysis of earlier sections, we can determine quickly which of these three outcomes will occur.

H. A Discrete Model

One of the fundamental assumptions of a differential equations model of an evolving arms race is that each nation can immediately and continuously change its expenditures in response to the current arms spending of both nations. Many people would argue that in the real world, countries set arms budgets perhaps once a year and can only change them at discrete intervals of time, not at every instant. It's instructive, then, to examine what a discrete version of Richardson's conceptualization of an arms race might look like.

The simplest translation from the continuous to the discrete would be to replace the derivative as the measure of rate of change with the actual difference in two successive time periods. Thus, if $B(i)$ and $R(i)$ represent the annual arms expenditures in Year I for Blue and Red, respectively, we would have

$$B(i + 1) - B(i) = a\,R(i) - m\,B(i) + r \tag{32}$$

and

$$R(i+1) - R(i) = bB(i) - n\,R(i) + s \tag{33}$$

for some positive constants a, b, m, and n and some constants r and s.

This system of two difference equations, which is an example of a *discrete dynamical system*, is the discrete Richardson arms race model. If we rewrite the system of equations in the form

$$\begin{aligned} B(i+1) &= B(i) + [a\,R(i) - m\,B(i) + r] \\ R(i+1) &= R(i) + [b\,B(i) - n\,R(i) + s] \end{aligned}$$

then it easy to see that this system is mathematically the same as the one produced by the Euler method,

$$\begin{cases} x_{i+1} = x_i + (ay_i - mx_i + r)\Delta t \\ y_{i+1} = y_i + (bx_i - ny_i + s)\Delta t \end{cases}$$

with $x_i = B(i)$, $y_i = R(i)$ and $\Delta t = 1$.

Example

Examine the discrete Richardson model,

$$\begin{aligned} B(i+1) - B(i) &= 0.5\,R(i) - 0.9\,B(i) + 0.3 \\ R(i+1) - R(i) &= 0.7\,B(i) - 0.4\,R(i) + 0.2 \end{aligned}$$

with $B(0) = 15$ and $R(0) = 16$.

Iteration of the difference equations for five steps yields

i	*B*(*i*)	*R*(*i*)
0	15.00	16.00
1	9.80	20.30
2	11.43	19.24
3	11.06	19.75
4	11.28	19.79
5	11.32	19.97

Note that there is some fluctuation in Blue's values: initially at 15, it first decreases to 9.8 after 1 year, then increases to 11.43 after 2 years, drops to 11.06 at the end of 3 years, then increases to 11.28 and 11.32 in the fourth and fifth years. If we reiterate the system for a

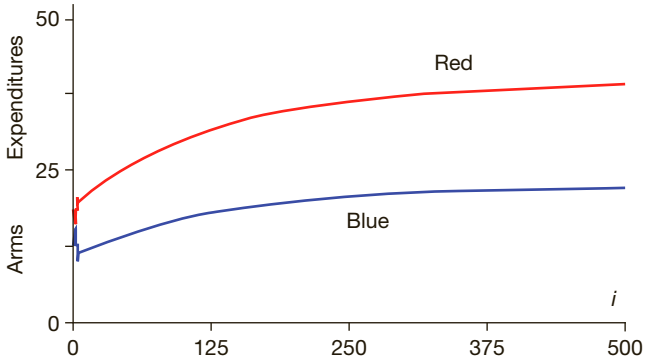


FIGURE 2.14 The stable point for this system occurs at (B*, R*) = (22, 39) where $0.5R^* - 0.9B^* + 0.3 = 0$ and $0.7B^* - 0.4R^* + 0.2 = 0$.

longer period—say, 400 years—then Blue's and Red's expenditures seem to reach stability. The values for i between 490 and 500 look like

I	*B*(*i*)	*R*(*i*)
490	21.753520	38.560150
491	21.755427	38.563554
492	21.757319	38.566931
493	21.759197	38.570282
494	21.761061	38.573608
495	21.762910	38.576907
496	21.764745	38.580181
497	21.766565	38.583430
498	21.768371	38.586653
499	21.770164	38.589852
500	21.771942	38.593026

A graph of the Blue and Red arms expenditures for the first 500 periods shows this stabilization more clearly. See Fig. 2.14.

V. Interpreting and Testing the Richardson Model

A. Interpretation

The Richardson model is simple and limited. It assumes that the rate of growth of armaments is influenced by only three factors and that these influences are additive in their effect. When you consider that many other forces operating in the real world may have an effect on arms races and that the interrelationships among these forces are undoubtedly quite complex, you might easily conclude that the Richardson model is too simplified to be of any real interest.

On the other hand, we have seen that the model does include—in mathematical language—some of the most common arguments about arms races that have been made by political analysts. Moreover, when we examine the mathematical implications of the model,

we see that most of them are in accord with common sense. The conclusions of our mathematical analysis in Section IV can be interpreted to give us these qualitative statements:

1. The presence of permanent underlying grievances will prevent total disarmament. As long as such grievances exist, countries will continue to arm even if their "rivals" have no weapons.
2. A stabilization of the arms race is achievable if the amount of mutual fear (measured by ab) is sufficiently tempered by the constraints (m and n) on the sizes of armament budgets.
3. Total disarmament is possible if there are underlying feelings of goodwill, but this will not occur if the level of armament expenditures is already above a certain critical amount or if mutual fear is too strong.

In addition to this qualitative information, the Richardson model also predicts how the arms race will develop quantitatively over time. In the case of a stabilized arms race, for example, the model tells what path will be taken toward the stable point.

If we wish to test this theoretical model against a real arms race, our first task is to consider more carefully how to measure the variables x and y. Next we would need to determine how to assign weights to the six parameters, or at least how to assess the signs of r and s and the relative magnitudes of mn and ab.

It would be useful at this point to discuss the history of the Richardson model and how it came to be derived in order to explain the cause of a world war.

B. Lewis Fry Richardson

The mathematical model of an arms race that we have been studying was the creation of a man named Lewis Fry Richardson. Richardson was born on October 11, 1881, at Newcastle-upon-Tyne in the county of Durham, England. His father had a tanning business, and his mother came from a family of corn merchants. Richardson attended Cambridge University where he studied under the famous Cavendish Professor of Physics Sir J. J. Thompson, discoverer of the electron. Richardson worked variously as a chemist, physicist, meteorologist, teacher of physics, and president of a technical college.

Photograph reproduced by permission of Stephen A. Richardson

Lewis F. Richardson

Richardson's scientific work in the field of meteorology was highly regarded. His book, *Weather Prediction by Numerical Process*, was published in 1923 and is considered a classic work in the field. The excellence of Richardson's contributions to meteorology and physics journals led to his receiving a Doctor of Science degree from London University in 1926, and to his election to the Fellowship of the Royal Society the next year. In 1972, Prime Minister Edward Heath opened the Richardson Wing, a major extension of the British Metrological Office's headquarters, named in Lewis's honor.

Richardson's family had a strong attachment to the Quaker religious community and the Society of Friends was a persistent influence in his life. He once wrote, "Its solemn emphasis on public and private duty . . . its condemnation of war pulled me away from the many warlike applications of physics." Richardson served with an ambulance convoy attached to the 16th infantry division of the French army during World War I. It was during this period that he began to write about the causes and avoidance of war. His short book *The Mathematical Psychology of War*, Oxford: Hunt, 1919, begins with a model embodying the mutual fear component of arms races.

Richardson had left his post as superintendent of the Eskdalemuir Observatory in Scotland, where he began his research on weather prediction, to join the Friends Ambulance Unit. After the war, he returned to his research at the Meteorological Office's Benson Observatory, but resigned in the summer of 1920 when the Office was put under the direction of the Air Ministry; his Quaker beliefs would not permit him to work directly for the armed services. His wife Dorothy Garnett Richardson later recalled, "there came a time of heartbreak when those most interested in his 'upper air' researches proved to be the 'poison gas' experts. [He] stopped his meteorological researches, destroying those that had not yet been published. What this cost him none will ever know."

Richardson developed his model of a two-nation arms race during the middle 1930s. He later extended the model to describe an arms race among n nations and tried to apply this refined model to the situation in Europe. He submitted a paper on this to an American journal urging immediate acceptance because he thought its publication might avert an impending war. The editors rejected the paper.

Not long after the outbreak of World War II, Richardson resigned his principalship of the Technical College and School of Art in Paisley, Scotland. In his retirement he continued to pursue his researches into the causes of war. In the last few years of his life, he returned to his earlier interests in meteorology. Richardson died on September 30, 1953.

In a 1953 obituary about Richardson, the meteorologist Peter A. Sheppard cited his books on weather prediction and international relations:

> *These mentally adventurous works stamp the man, and of him it may be truly said, as Wordsworth said of Newton '. . . a mind for ever/voyaging through strange seas of Thought, alone.' For it was given to Richardson to be way out ahead of his contemporaries in the effort to mould experience to scientific form. Some have now caught up, or are catching up, with the meteorological research which Richardson did thirty years ago; the fate of his pioneering efforts to form a science of international relations will perhaps not be known for a still longer time.*

Richardson's work also sowed the seeds for the applications of the exciting current field of *fractals*. He was the first person to investigate the relation between the scale used to

measure an irregular curve—such as coastline, a rugged mountain border frontier, or a meandering river—and the resulting estimate of its length.

In his study of "contiguity," Richardson observed that Spain reported that its boundary with Portugal was 613 miles long, while Portugal claimed that the border's length was 754 miles, a difference of 23%. Similarly, the Netherlands and Belgium asserted different lengths for their common frontier: 236 miles according to one country and 279 miles according to the other, an 18% gap. He speculated that much of the discrepancy was due to a difference in length of the measuring instruments.

The prototypical example for a fractal is the length of a coastline measured with different length rulers. Coastlines are very irregular and winding, characterized by bays, peninsulas, and inlets. Shorter measuring sticks fit more snugly in these bends and increase the estimated total length. If we use a yardstick to measure the length of Britain's coastline, we will get a smaller value than if we use a foot-long ruler.

One result of Richardson's studies was a graph showing the relation between scale and length for a variety of coasts, using logarithmic scales so that the exponent in the length formula can be read from the slope of a line fitted to the data points. Richardson's empirical studies suggested that for each nation's coastline, there are constants λ and D, such that if a ruler of length r was used, then one will obtain roughly λr^{-D} intervals, each of length r. The length of the coastline would be proportional to $\lambda\, r^{-D} r = \lambda\, r^{1-D}$.

Richardson's work proved to be the inspiration for Benoit Mandelbrot's suggestion that even though the exponent D was not a whole number, it should be regarded as a dimension. He termed it a "fractional dimension" or "fractal." Mandelbrot's book *The Fractal Geometry of Nature* uses the length of the coastline of Britain measured at various scales to introduce the notion of fractal dimension.

According to Mandelbrot (1924−2010), a *fractal* is an object or quantity that displays self-similarity, in a somewhat technical sense, on all scales. The object need not exhibit exactly the same structure at all scales, but the same "type" of structures must appear on all scales. A plot of the quantity on a log-log graph versus scale then gives a straight line whose slope is said to be the *fractal dimension*.

Many objects in nature are so complicated and irregular that they cannot be modeled well using familiar objects of classical geometry. "Clouds are not spheres, mountains are not cones, coastlines are not circles, and bark is not smooth, nor does lightning travel in a straight line," Mandelbrot observed He conceived and developed a new geometry, the geometry of fractal shapes, to model nature more accurately. Fractals are now used to model a wide range of biological and topographical entities and to produce ultrarealistic special effects for movies and video games.

Mandelbrot stated that Richardson's contiguity paper was a revelation for him and had profoundly affected his research. In *The Fractal Geometry of Nature*, he describes Richardson as a "great scientist" and notes, "[W]e are indebted to him for some of the most profound and most durable ideas regarding the nature of turbulence, notably the notion that turbulence involves a self-similar cascade." In a brief biographical sketch, Mandelbrot quotes Richardson on turbulence and self-similarity:

> *Big whorls have little whorls,*
> *Which feed on their velocity;*
> *And little whorls have lesser whorls,*
> *And so on to viscosity.*

C. Background of Richardson's Model

Manifestations of hostility among nations are often apparent in the invectives that appear in speeches and in the press. The difficulty of incorporating hostility as a variable in a mathematical model is in finding a suitable, objectively measurable quantity to represent the amount of hostility.

Richardson saw armament expenditures in monetary units as a good index of hostility. He proposed that indices of international trade be used as measures of the amount of cooperation between nations. Accordingly, the net amount of hostility would be the difference between arms expenditures and international trade. If this quantity is negative, then the magnitude of difference could be interpreted as net cooperation. Note that this would enable us to attach meaning to negative values of x and y in the model.

In developing his model, Richardson was attempting to discover the causes of World War I. He assumed that when armaments can reach constant equilibrium values, then no war occurs. If the armaments increase indefinitely, he concluded that war would eventually start.

Richardson [1960b, 15] defended his inclusion of what we have called the "mutual fear" factor in his model by quoting Sir Edward Grey, who was the British Foreign Secretary at the outbreak of World War I:

> *The increase of armaments that is intended in each nation to produce consciousness of strength, and a sense of security, does not produce these effects. On the contrary, it produces a consciousness of the strength of other nations and a sense of fear. . . . The enormous growth of armaments in Europe, the sense of insecurity and fear caused by them—it was these that made war inevitable. . . . This is the real and final account of the origin of the Great War.*

As to the presence of terms involving the burdens of arms expenditures, Richardson notes the remarks of Winston Churchill and Prince Bernhard von Bülow. Churchill records that on November 3, 1909, when he was president of the Board of Trade, he began a memo to the British cabinet with these words:

> *Believing that there are practically no checks upon German naval expansion except those imposed by the increasing difficulties of getting money, I have had the enclosed report prepared with a view to showing how far those limitations are becoming effective. It is clear that they are becoming terribly effective.*

Prince von Bülow, who was the German Chancellor, wrote in 1914 [Richardson, 1960b, 15]:

> *It is just possible that the effect of convulsively straining her military resources to the uttermost may, by reacting on the economic and social conditions of France, hasten the return of pacific feelings. . . . Should the 3-year military service entail an income tax, this would also probably have a sobering effect.*

In 1935, when Grey's statement that the enormous growth of armaments was the real cause of the war was quoted in a Parliamentary debate, L. S. Amery, said in reply [Richardson, 1960b, 15–16]:

> *With all respect to the memory of an eminent statesman, I believe that statement to be entirely mistaken. The armaments were only the symptoms of the conflict of ambition and ideals, of*

those nationalist forces, which created the war. . . . It was insoluble conflicts of ambitions and not in the armaments themselves that the cause of the war lay.

It was statements like Amery's that impelled Richardson to include the terms r and s measuring underlying grievances into his system of differential equations.

Thus, we see how Richardson was led to include the three causes of armament expenditure increases and decreases into his model, and how he arrived at armament expenditures as an indication of hostility. Is there some way to test this model? What predictions can we make from it that can be compared to reality?

D. Testing the Model

In this section, we will test the Richardson model against the actual arms race that took place in Europe in the years prior to the outbreak of World War I.

In the first decade of this century, it was apparent to many observers that there was a great likelihood of a war arising. The principal foes in the war would be France and Germany. It was clear that France would be allied with Russia and Germany with Austria-Hungary. It was thought that Great Britain would most likely support France and Russia, but the role of some other important European nations (Italy, Turkey) was in doubt.

Richardson attempted to test his model as an arms race between two blocs: France and Russia on one side, Germany and Austria-Hungary on the other. He began with assumptions that the degree of "mutual fear" and the braking effects of high armament budgets were the same on both sides—that is, he set $a = b$ and $m = n$ in Eqs. (11) and (12) to obtain

$$\frac{dx}{dt} = ay - mx + r \tag{34}$$

$$\frac{dy}{dt} = ax - my + s \tag{35}$$

If we add these equations, we obtain

$$\frac{d(x+y)}{dt} = (a-m)(x+y) + (r+s) \tag{36}$$

Finally, if we set $z = x + y$, we obtain the differential equation

$$\frac{dz}{dt} = (a-m)\left(z + \frac{r+s}{a-m}\right) \tag{37}$$

The variable z represents the total arms expenditures for both sides. Eq. (37) then makes a prediction that can be checked: *Total armament expenditures will increase at a rate that is proportional to total expenditures*. Mathematically, this asserts that if we plot dz/dt against z, we should obtain a straight line.

Richardson tabulated the armament budgets in millions of pounds sterling of the four powers in the years immediately before World War I. He estimated dz/dt for each 2-year period by simply taking the differences in total budgets for the years and then plotted this

Table 2.1

	1909	1910	1911	1912	1913
France	48.6	50.9	57.1	63.2	74.7
Russia	66.7	68.5	70.7	81.8	92.9
Germany	20.8	23.4	23.4	25.5	26.9
Austria-Hungary	199.2	204.8	214.9	238.7	289.0
Totals		5.6	10.1	23.8	50.3
Increases		202.0	209.8	226.8	263.8
2-year average					

Source: Adapted from Richardson, *Arms and Insecurity*, 1960, 32.

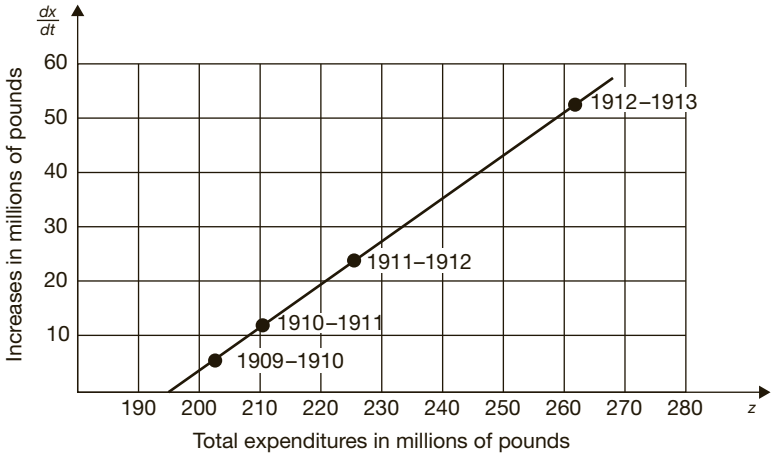


FIGURE 2.15 Rate of growth of z versus z. (Adapted from Richardson, 1960b, 33.)

difference against the average total for the 2 years. The data is presented in Table 2.1, and the graph of his results in Fig. 2.12.

The four points do lie close to a straight line. Richardson [1960b, 33] himself wrote:

> *Since I first drew this diagram . . . I have been incredulous about the marvelously good fit. Yet there is certainly no simple mistake. . . . The regularity of these phenomena shows that foreign politics had then a rather machine-like quality, intermediate between the predictability of the moon and the freedom of an unmarried young man.*

The slope of the line in Fig. 2.14 is about 0.73. The slope predicted by the model is $a - m$. Richardson gives some data to indicate that a reasonable figure for m is .2 so that a is roughly .9. Since $a = b$ and $m = n$ in this situation, we have $ab = .81$ and $mn = .04$. Since $ab > mn$, we have the ambiguous case.

If we extrapolate the observations in Fig. 2.14 along the straight line to the point where $dz/dt = 0$, we find that there $z = 194$ million pounds sterling. At this level of expenditures, the total expenditures would remain constant. Richardson [1960b, 33–34] concludes

> *As love covereth a multitude of sins, so the good will between opposing alliances would just have covered 194 million pounds of defense expenditure on the part of the four nations concerned. Their actual expenditure in 1909 was 199 million; and so began an arms race which led to World War I.*

The critical reader may not be so quick to accept this argument as evidence that the Richardson model is an accurate description of the dynamics of arms races. We have already pointed out that a number of simplifying assumptions were made in the original formulation of the model. More simplifying suppositions were necessary to make the prediction Richardson claims is verified by the facts. You might well consider, for example, why Richardson left out the armaments of Great Britain in his calculations when it was evident that this nation was an ally of France and Russia. Does the model justify the statement that a stabilized arms race implies that there will be no wars?

In an extended survey, *Civilizations*, *Empires and Wars*: *A Quantitative History of War*, William Eckhardt concluded, "The positive correlations between war preparations and the frequency and intensity of war convinced all authors that preparations provoked war more than they deterred them, thus confirming the arms race theory of war rather than the deterrence theory."

Attempts to validate Richardson's model with other real-world data have had mixed results, in part because accurate data is difficult to collect. Evidence from the most celebrated arms race of the 20th century—the nuclear buildups of the United States and the Soviet Union—indicates that the United States did react to perceptions of escalations of arms on the part of the Soviets, but insufficient data is available to determine whether the leaders in Moscow based their changes in expenses on what they thought Washington was spending. The Richardson model does seem to capture some of the dynamics of the continuing arms race between India and Pakistan and between Israel and its Arab neighbors, but fails to match the observations of an arms race between Greece and Turkey. The papers by Etcheson, Intrilligator and Brito, and by Isard and Anderson listed in the References provide excellent reviews of arms race models and their relevance.

We have presented Richardson's model not because we believe in all its assumptions or in its universal applicability to all arms races, but rather to illustrate model building, improving, interpreting, and testing procedures. In commenting on the importance of Richardson's work, Rashevsky and Trucco [Richardson, 1960b, Preface p. ix] state quite well the case for studying such a model:

> *The value of this work is not in the particular formulation of his theory but in the fact that Richardson shows how the problems of the causes of war can be subject to mathematical treatment and to rigorous mathematical thought. Even in physics, no matter how good a mathematical theory of a given set of phenomena is, it is eventually improved almost to a point beyond recognition. But the basic ideas of a good theory remain through all those changes. Look at the difference between Planck's original formulation of discontinuous emission of radiation and the present-day formulation of quantum mechanics. Yet the latter would not have been possible without the former. Richardson's equations will be changed by future investigators, some of his conclusions will be abandoned, but his work will remain forever as the first study of war on a rigorous basis of mathematical reasoning. Whatever the shortcomings of this model, it will have to be studied by every investigator who delves into the causes and origins of war. This work is a starting point for the development of new branch of sociology.*

The practical aspects of any theory seldom come at once. Radio, which was made possible by Maxwell's theory, came long after his death. Einstein's theoretical prediction of the equivalence of mass and energy remained for 40 years without any applications. When they came, they came with a vengeance. Richardson may have overemphasized the immediate applicability of his work. Its long-range usefulness cannot be doubted.

VI. Obtaining an Exact Solution

The Richardson arms race model is, mathematically speaking, a linked system of two first-order linear differential equations with constant coefficients. It is possible to find explicit solutions to such a system using the tools of calculus. We outline in this section how to accomplish this task.

We begin with a simpler first-order differential equation,

$$x' + cx = 0 \tag{38}$$

where x is a positive-valued function of t, c is a constant, and $x' = x'(t) = \dfrac{dx}{dt}$

Observe that we can rewrite equation (38) as

$$x' = -cx \quad \text{or} \quad \frac{1}{x}x' = -c \tag{39}$$

If we integrate both sides of this equation with respect to t, we obtain

$$\int \frac{1}{x}x'\,dt = \int -c\,dt$$

or

$$\ln x = -ct + A$$

which we can write as

$$x = Be^{-ct}$$

for an arbitrary constant B.

Now let's consider a more complicated second-order differential equation,

$$x'' + bx' + cx = 0 \tag{40}$$

where b and c are constants. Could this equation also have a solution of the form $x = B\,e^{ut}$ for some constants B and u?

Let's try substituting such a function into the differential equation, noting that $x' = Bue^{ut}$ and $x'' = B\,u^2e^{ut}$. We obtain

$$B\,u^2\,e^{ut} + bBue^{ut} + cBe^{ut} = 0 \tag{41}$$

Since the exponential function never has value 0, this last equation is equivalent to

$$u^2 + bu + c = 0 \tag{42}$$

for any nonzero constant B.

Thus, we see that $x = A\,e^{ut}$ is a solution of $x'' + bx' + cx = 0$ if and only if u is a root of the quadratic equation $u^2 + bu + c = 0$.

What does all this have to do with the Richardson arms race model? Let's consider a particular example:

$$\begin{aligned} dx/dt &= y - 6x + 16 \\ dy/dt &= 2x - 5y + 4 \end{aligned} \tag{43}$$

which has stable point (3, 2).

We first make the change of variable $X = x - 3$, $Y = y - 2$.
Then

$$X' = x' = y - 6x + 16 = (Y + 2) - 6(X + 3) + 16 = Y - 6X + 2 - 18 + 16 = Y - 6X.$$

Similarly, $Y' = 2X - 5Y$.

If we can solve the simpler system

$$\begin{aligned} X' &= Y - 6X \\ Y' &= 2X - 5Y \end{aligned} \tag{44}$$

for X and Y as explicit functions of t, say $X = f(t)$ and $Y = g(t)$, then we can obtain solutions to the original system as

$$x = X + 3 = f(t) + 3$$

and (45)

$$y = Y + 2 = g(t) + 2.$$

To obtain solutions for the simpler system (Eqs. 44), note that the first equation gives us $Y = X' + 6X$, so we can express the second equation as

$$Y' = 2X - 5(X' + 6X) = -5X' - 28X$$

Now if we differentiate the first equation of (44) with respect to t, we obtain

$$\begin{aligned} X'' &= (Y - 6X)' = Y' - 6X' \\ &= -5X' - 28X - 6X' = -11X' - 28X \end{aligned}$$

or

$$X'' + 11X' + 28X = 0 \tag{46}$$

From our work above, we see that that $X = A\,e^{ut}$ is a solution to $X'' + 11X' + 28X = 0$ when u is a root of $u^2 + 11u + 28 = (u + 4)(u + 7) = 0$—that is, u equals -4 or -7.

Hence, Ae^{-4t} and Be^{-7t} are each solutions of $X'' + 11X' + 28X = 0$ for any constants A and B. It is easy to see, by substitution into the differential equation, that $X = f(t) = Ae^{-4t} + Be^{-7t}$ is a solution also for any constants A and B.

It is a standard theorem proved in differential equations or linear algebra courses that every solution of $X'' + 11X' + 28X = 0$ *must* be of the form $X = f(t) = Ae^{-4t} + Be^{-7t}$ for some constants A and B. The interested reader may wish to consult the book by Sanchez or the one by Brauer, Nohel, and Schneider listed in the References.

Since $Y = X' + 6X$, we have

$$\begin{aligned} Y = g(t) &= f'(t) + 6f(t) \\ &= -4Ae^{-4t} - 7Be^{-7t} + 6(Ae^{-4t} + Be^{-7t}) \\ &= 2Ae^{-4t} - Be^{-7t} \end{aligned}$$

and hence the solution of our original system

$$\begin{aligned} dx/dt &= y - 6x + 16 \\ dy/dt &= 2x - 5y + 4 \end{aligned}$$

is

$$\begin{aligned} x(t) &= Ae^{-4t} + Be^{-7t} + 3 \\ y(t) &= 2Ae^{-4t} - Be^{-7t} + 2 \end{aligned}$$

where the constants A and B depend on the initial state of the arms race.

Observe that since $\lim_{t\to\infty} e^{-4t} = 0 = \lim_{t\to\infty} e^{-7t}$, we have

$$\lim_{t\to\infty} x(t) = 3 \text{ and } \lim_{t\to\infty} y(t) = 2$$

so this is a stable arms race.

Let's turn now to the general Richardson arms race model:

$$\begin{aligned} dx/dt &= ay - mx + r \\ dy/dt &= bx - ny + s \end{aligned} \tag{47}$$

and undertake a similar analysis.

First, we make the change of variable

$$\begin{aligned} X &= x - x^* \\ Y &= y - y* \end{aligned} \tag{48}$$

where (x^*, y^*) are the coordinates of the stable point. Our system of differential equations takes the form

$$\begin{aligned} X' &= aY - mX \\ Y' &= bX - nY \end{aligned} \tag{49}$$

The first equation gives us $aY = X' + mX$ so that $Y = \frac{x'}{a} + \frac{mX}{a}$ and since $Y' = bX - nY$, we have

$$Y' = bX - \frac{n}{a}X' - \frac{mn}{a}X = \frac{ab - mn}{a}X - \frac{n}{a}X'$$

Differentiating the first equation of (49) yields

$$X'' = aY' - mX' = (ab - mn)\,X - nX' - mX'$$

so

$$X'' + (m+n)X' + (mn - ab)X = 0 \tag{50}$$

The solutions of this differential equation will come, as we've seen above, from the roots of the quadratic equation:

$$u^2 + (m+n)u + (mn - ab) = 0 \tag{51}$$

which are

$$u = \frac{-(m+n) \pm \sqrt{(m+n)^2 - 4(mn - ab)}}{2} \tag{52}$$

and these can also be expressed as

$$u = \frac{-(m+n) \pm \sqrt{(m-n)^2 + 4ab}}{2} \tag{53}$$

Since a and b are positive, the discriminant $(m-n)^2 + 4ab$ is also positive. We see from Eq. (53) that both roots are real. Furthermore, the root

$$u_1 = \frac{-(m+n) - \sqrt{(m+n)^2 - 4(mn - ab)}}{2}$$

is negative. If, in addition, $mn > ab$, then $(m+n)^2 - 4(mn - ab)$ will be less than $(m+n)^2$, so the second root $u_2 = \frac{-(m+n) + \sqrt{(m+n)^2 - 4(mn - ab)}}{2}$ would also be negative. On the other hand, if $mn < ab$, then the root u_2 will be positive.

The solution of Eqs. (49) will have the form

$$X = f(t) = Ae^{u_1 t} + Be^{u_2 t}$$

for some constants A and B, and the solution for $x(t)$ in the original arms model will be

$$x = Ae^{u_1 t} + Be^{u_2 t} + x^* \tag{54}$$

From $aY = X' + mX$, we have $aY = Au_1 e^{u_1 t} + Bu_2 e^{u_2 t} + mAe^{u_1 t} + mBe^{u_2 t}$

Or, collecting terms, $aY = A(u_1 + m)e^{u_1 t} + B(u_2 + m)e^{u_2 t}$. Hence, we will have as the solution for $y(t)$

$$y = \frac{A(u_1 + m)e^{u_1 t} + B(u_2 + m)e^{u_2 t}}{a} + y^* \tag{55}$$

If both u_1 and u_2 are negative, then x approaches x^* and y approaches y^* as time progresses; the arms race will be stable. If u_2 is positive, then the eventual outcome of the arms race depends on the sign of B. If B is positive, then we will have a runaway escalating race, while if B is negative, the race will be toward mutual disarmament.

The sign of B is dependent on the initial values of x and y. The following example illustrates this dependence.

Example

The arms race governed by the equations:

$$dx/dt = -9x + 11y - 15$$
$$dy/dt = 12x - 8y - 60$$

has stable point $(x^*, y^*) = (13, 12)$ so the change of variable $X = x - 13$, $Y = y - 12$ converts the system to

$$dX/dt = -9X + 11Y$$
$$dY/dt = 12X - 8Y$$

and this system yields the second-order differential equation:

$$X'' + 17X' - 60 = 0.$$

The corresponding quadratic equation is $u^2 + 17u - 60 = (u + 20)(u - 3) = 0$, which yields the solution:

$$X = f(t) = Ae^{-20t} + Be^{3t}$$

$$Y = g(t) = \frac{X' + 9X}{11} = \frac{-20A\,\mathrm{e}^{-20t} + 3B\,\mathrm{e}^{3t} + 9A\,\mathrm{e}^{-20t} + 9B\,\mathrm{e}^{3t}}{11} = -Ae^{-20t} + \frac{12}{11}Be^{3t}$$

and so the solutions of the original model are

$$x = Ae^{-20t} + B\,\mathrm{e}^{3t} + 13$$
$$y = -Ae^{-20t} + \frac{12}{11}B\,\mathrm{e}^{3t} + 12$$

The ultimate behavior of this arms race depends on the sign of B, which in turn hinges on the initial values

$$x_0 = x(0) = A + B + 13 \text{ and } y_0 = y(0) = -A + \frac{12}{11}B + 12.$$

The solution of this pair of equations is

$$A = \frac{-24 + 12x_0 + 11y_0}{23}$$

$$B = \frac{-275 + 11(x_0 + y_0)}{23}.$$

Now B is positive when $11(x_0 + y_0) > 275 = (11)(25)$—that is, $(x_0 + y_0) > 25$. Hence this arms race will result in runaway escalation if the sum of the initial expenditures exceeds 25, but will move toward mutual disarmament if that sum is less than 25.

Fig. 2.16 shows the graphs of $x(t)$ and $y(t)$ in a case where $(x_0 + y_0) = 24.5$. Fig. 2.16 displays a graph of the orbit in the (x, y)-plane.

Finally, we note that matrix algebra (see Appendix II) provides another way to find the values of u_1 and u_2. If A is the 2-by-2 matrix

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$$

then the *determinant* of A, denoted *det* A, is the number defined by

$$det\, A = a_{11}\, a_{22} - a_{12}\, a_{21}$$

The *eigenvalues* of A are the roots of the quadratic equation $det\, B = 0$, where B is the matrix

$$B = A - uI = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} - \begin{pmatrix} u & 0 \\ 0 & u \end{pmatrix} = \begin{pmatrix} a_{11} - u & a_{12} \\ a_{21} & a_{22} - u \end{pmatrix}$$

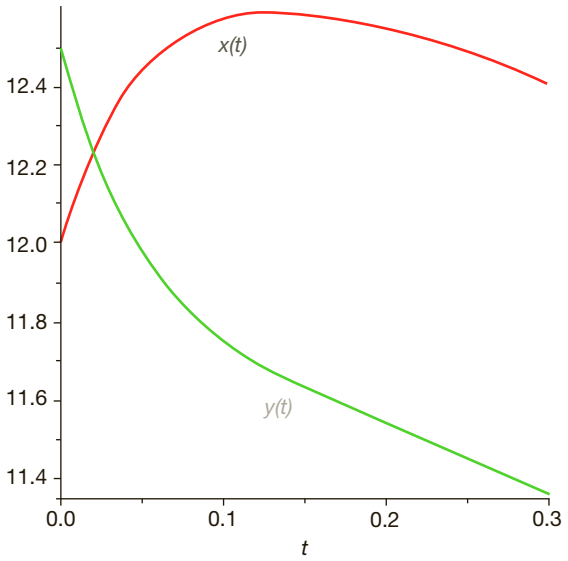


FIGURE 2.16 Graphs of the solutions of the arms race $dx/dt = -9x + 11y - 15$, $dy/dt = 12x - 8y - 60$ for initial values $x_0 = 12$; $y_0 = 12.5$.

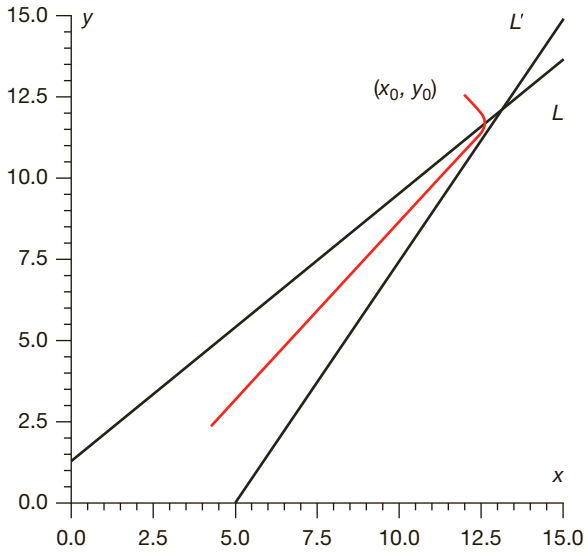


FIGURE 2.17 Graphs of the orbit of the arms race $dx/dt = -9x + 11y - 15$, $dy/dt = 12x - 8y - 60$ with initial values $x_0 = 12$; $y_0 = 12.5$ and of the stable lines L and L'.

Hence, if A is the matrix of coefficients of the system

$$X' = -mX + aY$$
$$Y' = bX - nY$$

which is to say

$$A = \begin{pmatrix} -m & a \\ b & -n \end{pmatrix}$$

then the eigenvalues of A are the roots of the quadratic equation

$$\det \begin{pmatrix} -m-u & a \\ b & -n-u \end{pmatrix} = 0$$

A direct computation shows that the eigenvalues of A are exactly the roots of the quadratic equation Eq. (51). See Exercise 45.

EXERCISES

I. The Real-World Setting

1. Do historians believe that "mutual fear" drove the escalation in nuclear armament of the United States and the Soviet Union during the Cold War?
2. What evidence can you find that a "mutual fear" situation exists today in the Middle East, in East Africa, between India and Pakistan, or between some other pair of belligerent nations?
3. Although Richardson's primary concern was arms races between nations, others have observed similar mutual escalation and de-escalation of weapons procurement within individual nations. Examples would be a government and internal rebels, rival urban street gangs, and law enforcement agencies versus organized crime. Can you suggest other possible situations in

which the assumptions of Richardson's models would be applicable?

II. Constructing a Deterministic Model

4. Comment on the importance of the assumption that the variables in a deterministic model must represent observable and measurable quantities.

III. A Simple Model for an Arms Race

5. Let $a = b = 1$ in Eqs. (1) and (2).

(a) Verify that $f(t) = Ce^t + De^{-t}$ and $g(t) = Ce^t - De^{-t}$ give a solution to the system of differential equations, where C and D are arbitrary constants.

(b) Find the appropriate values of C and D if $x_0 = 3$ and $y_0 = 1$.

(c) Sketch the graphs of f and g.

(d) Determine $\lim_{t\to\infty} f(t)$ and $\lim_{t\to\infty} g(t)$.

6. Show that the constant C of Eq. (8) is equal to $y_0^2 - (b/a)x_0^2$.

7. What happens in the simple arms race model in each of the following cases?

(a) $x_0 = y_0 = 0$.

(b) $x_0 = 0$ and y_0 is positive.

(c) x_0 is positive and $y_0 = 0$.

(d) $x_0 = y_0$ is positive.

8. Analyze the simple arms race model if a and b are both negative. What assumptions does this model reflect? Can these assumptions be defended?

9. Analyze the simple arms race model if a and b have opposite signs. What assumptions does this model reflect? Can these assumptions be defended?

10. Is it possible that $\big(x(t), y(t)\big)$ asymptotically approaches some point (x^*, y^*) in the first quadrant with x^* and y^* both positive as time goes on? Why?

11. This problem develops a closed form solution for the function $x = f(t)$ of the simple arms race model.

(a) Show that differentiation of Eq. (1) gives $x''(t) = ay'(t) = abx(t)$.

(b) By substituting $x = e^{mt}$ into the equation of part (a), show that there are two values for the constant m—call them m_1 and m_2—that make the equation valid.

(c) Let $f_1(t) = e^{m_1 t}$ and $f_2(t) = e^{m_2 t}$. Show that if $x(t) = Cf_1(t) + Df_2(t)$ where C and D are any constants, then $x''(t) = abx(t)$.

(d) If $x = f(t)$ and $y = g(t)$ are solutions satisfying Eqs. (3) and (4), show that $f(0) = x_0$ and $f'(0) = ay_0$ while $g(0) = y_0$ and $g'(0) = bx_0$.

(e) If the function f satisfies Eqs. (3) and (4) where $f(t) = Cf_1(t) + Df_2(t)$, show that $C + D = x_0$ and $Cm_1 + Dm_2 = ay_0$.

(f) Solve the equations of (e) for C and D in terms of x_0, y_0, m_1 and m_2.

(g) Show that if $f(t)$ satisfies Eqs. (3) and (4) and is given by (e), then

$$f(t) = \frac{(\sqrt{ab}\,x_0 + ay_0)e^{\sqrt{abt}} + (\sqrt{ab}\,x_0 - ay_0)e^{-\sqrt{abt}}}{2\sqrt{ab}}$$

(h) Let $f(t)$ be the function of part (g). Find $\lim_{t\to\infty} f(t)$.

(i) Sketch a graph of f.

12. Find $g(t)$ explicitly as a function of t so that g satisfies Eqs. (3) and (4). Carry out the steps analogous to parts (a)–(g) of Exercise 11.

IV. The Richardson Model

13. Show, by substitution into the differential equations, that $x = Ae^t + 4Be^{-6t} - \frac{5}{3}$, $y = Ae^t - 3Be^{-6t} - \frac{3}{2}$ is a solution of

$$\frac{dx}{dt} = 4y - 3x + 1$$

$$\frac{dy}{dt} = 3x - 2y + 2$$

for any choice of constants A and B. What is the ultimate behavior of an arms race with these equations? Sketch the graphs of x and y as functions of t. If the expenditures at $t = 0$ are x_0 and y_0, find A and B in terms of x_0 and y_0.

14. Discuss the effect of setting r and s both equal to 0 on the question of stability—that is, investigate the consequences of the model given by Eqs. (9) and (10).

15. Discuss the stability question for the Richardson model in the cases

(a) $a = 0$

(b) $n = 0$

Begin by determining the equations of the optimal lines.

16. Can the lines L and L' be parallel? What happens to the arms race in this case?

17. Show that it is in fact possible for L and L' to be identical! Find a specific set of values for a, b, m, n, r, and s for which this happens. What is the long-term behavior of an arms race if the lines L and L' coincide?

18. **(a)** Show that if $mn - ab \neq 0$, then the lines L and L' intersect at the point (x^*, y^*) where

$$x^* = \frac{rn + as}{mn - ab} \text{ and } y^* = \frac{br + ms}{mn - ab}$$

(b) If the grievance terms r and s are positive, show that the stable point (x^*, y^*) lies in the first or third quadrant of the plane, depending on the sign of $mn - ab$.

19. Prove that dy/dt is negative at every point above L' and positive at every point below L'. Show that this implies that Red always adjusts its expenditures toward its optimal line. You will have to include a careful definition of what *above* and *below* mean in this setting.

20. In the Richardson model, is it possible that arms expenditure levels will tend to approach some point $(x^\#, y^\#)$ in the positive first quadrant that is not the stable point? Explain.

21. Verify the details of the argument that if r and s are positive and $mn - ab$ is negative, then the arms race always leads to runaway expenditures regardless of the location of the initial point.

22. Show that the lines L and L' can assume the configuration of Fig. 2.9 only when at least one of the "grievance" terms is negative.

23. Show that if r and s are negative and (x^*, y^*) is in the third quadrant, then the lines L and L' split the first quadrant into three regions and that the arms race tends to total disarmament for an initial point in any of the three regions.

24. **(a)** Carry out the Euler procedure for the arms race example in Section IV, part D with initial level (15, 24). What is the outcome in this case?

(b) Verify the result in (a) by using the "point-slope method."

25. Consider the differential equation $dx/dt = 2 - \frac{x}{t}$ with $x(1) = 2$.

(a) Verify that $x = t + \frac{1}{t}$ is a solution. What is $x(2)$?

(b) Use the Euler method to solve the differential equation. With $\Delta t = .1$, what value does this assign to $x(2)$?

26. Apply Euler's method to the example of Section IV, part D with $\Delta t = .1$ and $\Delta t = 1$. What conclusions can you draw about the outcome of the arms race?

27. Does Eq. (26) always determine two straight lines? If so, find the equations of these lines.

28. The system $dx/dt = y - 2x - 5$, $dy/dt = 6x - 2y - 12$, first introduced at the beginning of Section IV, part D, has the solution

$$x(t) = Ae^{(-2+\sqrt{6})t} + Be^{(-2-\sqrt{6})t} + 11$$

$$y(t) = \sqrt{6}Ae^{(-2+\sqrt{6})t} - \sqrt{6}Be^{(-2-\sqrt{6})t} + 27$$

for any constants A and B.

(a) Verify this by substitution into the system of differential equations.

(b) Evaluate A and B if at $t = 0$ we have $x_0 = y_0 = 15$.

(c) For the values of A and B obtained in part (b), determine the limiting behavior of this arms race using the explicit formulas of (a). Is the answer consistent with that obtained in Section IV, part D?

29. Determine the outcome of an arms race governed by the Richardson model

$$dx/dt = 10y - 14x - 12$$

$$dy/dt = 8x - 4y - 24$$

if

(a) initial level is (4, 4)

(b) initial level is (13, 6)

30. In his book *Arms and Insecurity*, Richardson also constructed a "rivalry" model to reflect the assumption that a state is threatened not by the total amount of arms the opponent has, but rather by the discrepancy

between its opponent's arms and its own. The revised model takes the form

$$dx/dt = a(y - x) - mx + r \quad \text{where } a, b, m, n$$
$$dy/dt = b(x - y) - ny + s \quad \text{are positive constants}$$

Show that if r and s are positive, then such an arms race is always stable.

What can you conclude about the long-term behavior of this arms race if no assumptions are made about the signs of r and s? In particular, prove that there can never be a runaway arms race for this model.

31. Analyze the long-term behavior of the discrete Richardson model

$$B(i+1) - B(i) = 0.5\,R(i) - 0.9\,B(i) + 0.3$$
$$R(i+1) - R(i) = 0.7\,B(i) - 0.4\,R(i) + 0.2$$

if $B(0) = 40$ and $R(0) = 50$ by iteration for a large number of steps.

32. Analyze the long-term behavior of the discrete Richardson model

$$B(i+1) - B(i) = 0.1\,R(i) - 0.2\,B(i) - 0.5$$
$$R(i+1) - R(i) = 0.6\,B(i) - 0.2\,R(i) - 1.2$$

if

(a) $B(0) = 15$ and $R(0) = 15$

(b) $B(0) = 15$ and $R(0) = 30$

33. Suppose $a = b$ and $m = n$ and let $T(i)$ be the two nation total of arms expenditures in Year I. Show that $T(i+1) - T(i) = (a - m)T(i) + q$ where $q = r + s$. Use mathematical induction to conclude that

$$T(i) = (1 + a - m)^i \left[T(0) + \frac{c}{a - m}\right] - \frac{c}{a - m}$$

If $m > a$, show that the arms race is stable and $T(i)$ approaches $\dfrac{c}{m - a}$. Discuss the outcome if $m < a$.

V. Interpreting and Testing the Richardson Model

34. Richardson's model predicts that the rate of change of total expenditures, $z = x + y$, is a linear function of total expenditures. Can you construct a different model of an arms race that leads to the same conclusion?

VI. Obtaining an Exact Solution of the Richardson Model

35. Show that the change of variables $X = x - x^*$, $Y = y - y^*$ has the geometric effect of replacing the standard (x, y) rectangular coordinate system with a new (X, Y) rectangular coordinate system where the stable point of the original pair of differential equations now sits at the origin of the (X, Y) plane.

36. Show that if $b^2 = 4c$, then the quadratic equation $u^2 + bu + c = 0$ has a single solution, $u = -b/2$. In such a case show that in addition to the function $x = e^{ut}$, the differential equation $x'' + bx' + cx = 0$ also has solution $x = t\,e^{ut}$.

37. Show that if $b^2 < 4c$, then the quadratic equation $u^2 + bu + c = 0$ has complex roots, and in this case the functions $e^{\frac{-b}{2}t}\cos\left(\frac{\sqrt{4c - b^2}}{2}t\right)$ and $e^{\frac{-b}{2}t}\cos\left(\frac{\sqrt{4c - b^2}}{2}t\right)$ are each solutions of $x'' + bx' + cx = 0$.

38. Find and sketch the graphs of the exact solutions of each of the following arms race models:

(a) $\dfrac{dx}{dt} = -5x + 2y + 7$, $\dfrac{dy}{dt} = 2x - 3y + 6$; $x(0) = 3$, $y(0) = 5$

(b) $\dfrac{dx}{dt} = -1x + 4y + 7$, $\dfrac{dy}{dt} = 3x - 2y + 1$; $x(0) = 3$, $y(0) = 5$

(c) $\dfrac{dx}{dt} = -4x + \dfrac{7}{4}y - 7$, $\dfrac{dy}{dt} = 1x - 1y - 5$; $x(0) = 3$, $y(0) = 5$

39. Does the long-term qualitative behavior of any of the arms race models in (a)–(c) change if you alter the initial values $x(0)$ and/or $y(0)$?

40. Find and sketch the graphs of the exact solutions of the arms race

$$\frac{dx}{dt} = -6x + 10y - 28, \quad \frac{dy}{dt} = 9x - 5y - 58$$

in the following cases:

(a) $x(0) = 12$, $y(0) = 16$

(b) $x(0) = 12$, $y(0) = 4$

(c) $x(0) = 12$, $y(0) = 10$

41. Show that if the stable lines L and L' are parallel, then the values of u_1 and u_2 are $(m + n)$ and 0. What do the exact solutions of the arms race model look like in this case? What is the long-term behavior?

42. Show that the arms race of Exercise 40 ends in mutual disarmament if $9x(0) + 10y(0) < 28$ and is a runaway

arms race if $9x(0) + 10y(0) > 28$. What happens in the long term if $9x(0) + 10y(0) = 28$?

43. A first-order differential equation $dy/dx = F(x, y)$ is called *separable* if F can be written as a product of a function of x and a function of y—that is, $dy/dx = f(x)\,g(y)$. A separable differential equation can be solved by rewriting it in integral form as $\int \frac{1}{g(y)}\,dy = \int f(x)\,dx$ and carrying out the indicated integration. Solve the differential equation $dy/dx = \frac{x^2}{y}$ with initial condition $y(3) = 4$.

44. This problem examines a different technique for finding an implicit equation for the orbit of a Richardson arms race. Show that the change of variable $Y = VX$ produces $dY/dX = V + X\,dV/dX$ and hence transforms the differential equation $\frac{dY}{dX} = \frac{bX - nY}{-mX + aY}$ into a *separable* differential equation in X and V. Solving that differential equation and replacing V with Y/X gives an equation for the orbit in the (X,Y) plane. Carry out this process for the Richardson arms race model of Eqs. (9) and (10).

45. For those with a linear algebra background: for the generic arms race, show that the roots of the associated quadratic equation are the eigenvalues of the coefficient matrix $\begin{pmatrix} -m & a \\ b & -n \end{pmatrix}$.

46. (Linear Algebra): If $u \neq v$, show that the functions e^{ut} and e^{vt} form a linearly independent set.

47. (Linear Algebra): If $u = v$, show that the functions e^{ut} and te^{ut} form a linearly independent set.

48. Show that in the symmetric arms race ($a = b$ and $m = n$), the roots of the associated quadratic are $m \pm a$.

49. Find explicit solutions to the following Richardson models:

(a) $a = 4$ $b = 3$ $m = 3$ $n = 2$ $r = 1$ $s = 2$

(b) $a = 1$ $b = 6$ $m = 2$ $n = 2$ $r = -5$ $s = -12$

(c) $a = 6$ $b = 4$ $m = 5$ $n = 5$ $r = 2$ $s = 3$

SUGGESTED PROJECTS

1. One can argue that in the real world, a runaway arms race is impossible since there is an absolute limit to the amount any country can spend on arms: the gross national product minus some amount for survival. How might this idea be incorporated into the Richardson model? One approach would be to let x_M and y_M be the maximum amounts that Blue and Red, respectively, could spend on arms; these terms are sometimes referred to as the *carrying capacities*. We could then form the model

$$\begin{aligned} \frac{dx}{dt} &= \left(1 - \frac{x}{x_M}\right)(ay - mx + r) \\ \frac{dy}{dt} &= \left(1 - \frac{y}{y_M}\right)(bx - ny + s) \end{aligned} \tag{56}$$

Analyze such a model using the techniques developed in this chapter.

2. Extend the Richardson model to the situation of three nations. Derive a set of differential equations if the three are mutually fearful so that each one is spurred to arm by the expenditures of the other two; examine the stability question for this example. Also derive equations if two of the nations are close allies who are not threatened by the arms buildup of each other but are threatened by the expenditures of the third; discuss the possibilities for stability in this case.

3. The basic assumptions of our model require that a, b, m, and n be positive numbers. If negative values are assigned to these, the model would go in reverse: the armaments of the rival would act as a brake and one's own armaments as a spur. Investigate the stability of such an arms race. Can such a model be defended on the basis of real-world observations?

4. Suppose the underlying differential equations have the form

$$\frac{dx}{dt} = ay^2 - mx + r$$

$$\frac{dy}{dt} = bx^2 - ny + s$$

where a, b, m, and n are positive. Sketch the stability curves $dx/dt = 0$ and $dy/dt = 0$. How many stable points are there? Discuss the outcomes of such an arms race for various intersections of the stability curves.

5. A nation (Blue, for example) may not be spurred to arm so much by the absolute level of its enemy's expenditures y, but rather by how much the other side is exceeding

its stable level, y^*. Show that one way to model this assumption would have $\frac{dx}{dt} = a(y - y^*) - mx + r$. Suppose Red has the same motivation. What are the possibilities for long-term behavior of such a model?

6. During his research work in Bali with Margaret Mead, the anthropologist Gregory Bateson became interested in the factors that keep cultures together or drive them apart, Bateson described two basic forms of relationship between groups in a culture, "symmetrical" and "complementary." In a symmetrical relation the same behavior is exchanged: more of it in Red is answered by more of it in Blue. In complementary relations, opposite and mutually dependent behaviors are exchanged. Bateson was influenced by Richardson's arms race model. Investigate Bateson's theory and explore how the Richardson model can be adapted to formulate the theory in mathematical terms. This provides another good example of how essentially similar mathematical models can be used to investigate real-world problems that at first sight appear to have little to do with each other.

7. Derive explicit solutions for the discrete version of the Richardson model.

$$B(i+1) - B(i) = a\,R(i) - m\,B(i) + r$$
$$R(i+1) - R(i) = b\,B(i) - n\,R(i) + s$$

A key step will be to show that there are solutions to the modified system

$$B(i+1) - B(i) = a\,R(i) - m\,B(i)$$
$$R(i+1) - R(i) = b\,B(i) - n\,R(i)$$

which are of the form

$$B(k) = A\,\lambda^k \cdot R(k) = A^*\lambda^k$$

Where λ is a solution of the quadratic $\lambda^2 + (m+n)\lambda + (mn - ab) = 0$.

8. During World War I, F. W. Lanchester developed some mathematical models of combat. In one of these models, Lanchester assumes that there are two combat forces in battle against each other. He assumes that these are "conventional" forces that operate in the open, comparatively speaking, and that every member of a force is within the "kill" range of the enemy. He also assumes that as soon as the conventional force suffers a loss, fire is concentrated on the remaining combatants. Finally, he assumes that each side is reinforced at a constant rate.

Show that Lanchester's assumptions are incorporated in the model

$$dx/dt = -ay + m$$
$$dy/dt = -bx + n$$

where a, b, m, and n are positive constants, t represents time, and x and y are the sizes of the two opposing forces.

Solve the system of equations explicitly in the case in which there are no reinforcements—that is, $m = n = 0$.

What can you say about the long-term behavior of this system?

For an application of this model to the Battle of Iwo Jima in World War II, see Martin Braun, *Differential Equations and Their Applications*, 4th ed., New York: Springer, 1993.

How might you modify this model if one of the forces is a guerrilla force? See S. J. Deitchman, "A Lanchester Model of Guerrilla Warfare," *Operations Research* **10** (1962): 818–827, 1962.

You can find a listing of references and suggestions for additional reading on the book's website, www.wiley.com/college/olinick

Ecological Models: Single Species

CHAPTER 3

The fact that ecology is essentially a mathematical subject is becoming ever more widely accepted. Ecologists everywhere are attempting to formulate and solve their problems by mathematical reasoning.

—Evelyn C. Pielou

I. Introduction

This chapter initiates the study of simple deterministic models for population growth. As Evelyn Pielou notes in her book *An Introduction to Mathematical Ecology*, "The investigation of the growth and decline of population is, historically, the oldest branch of mathematical ecology." Chapter 3 examines models for the changes in single-species population. The mathematical tools employed are first-order differential equations and first-order difference equations. In Chapter 4, we consider some models for population growth that present important features of interaction between two species occupying the same territory. In particular, we study the oscillation of population sizes of two competing species and the dynamics of predator-prey populations. Here the mathematical tool is an autonomous system of first-order differential equations. Chapter 5 presents some models on the growth of a population of cells making up a tumor. The mathematical analysis is self-contained.

II. The Pure Birth Process

Imagine a population made up entirely of identical organisms that reproduce at a rate that is the same for every individual and that does not vary with time. If we assume that each individual lives forever, that the organisms do not interfere with one another, and that there are sufficient space and resources to sustain all the individuals, then we are dealing with what ecologists term a "pure birth process." This process has been used to study yeast cells growing by fission, the propagation of new ideas, and the increase in the number of scientists over time, as well as many other types of population growth. The mathematical model for this process is a first-order differential equation,

$$\frac{dP}{dt} = bP \quad \text{or} \quad P'(t) = bP(t) \tag{1}$$

where $P = P(t)$ is the population at time t and b is the positive constant *birth rate* for each individual. We may also write this differential equation as

$$P'(t) = bP(t) \tag{2}$$

The opposite of the pure birth process—the *pure death process*—is described by essentially the same mathematical model. In the pure death process, we assume that no births occur and that each individual has the same positive likelihood d of death at every moment, a probability that does not change with time or with the age of the individual. The constant d is called the *death rate*. The differential equation describing these assumptions has the form

$$\frac{dP}{dt} = -dP \quad \text{or} \quad P'(t) = -dP(t) \tag{3}$$

The original model can also be employed to describe a population in which both births and deaths occur. Assume again that the birth rate b and the death rate d are positive constants independent of time, size of population, and age of individual. The model is the differential equation

$$\frac{dP}{dt} = (b-d)P \quad \text{or} \quad P'(t) = (b-d)P(t) \tag{4}$$

Setting $a = b - d$, we see that the same equation,

$$\frac{dP}{dt} = aP \quad \text{or} \quad P'(t) = aP(t) \tag{5}$$

describes all three situations.

Analysis of the model What are the mathematical consequences of this model, and what are the corresponding interpretations? First, "separate the variables" P and t to obtain

$$\int \frac{1}{P}\,dP = \int a\,dt \tag{6}$$

(More carefully, write Eq. (5) as $\frac{P'(t)}{P(t)} = a$ and integrate both sides with respect to t. Students who have not worked with differential equations before may wish to consult Appendix V.) Carry out the indicated integration to arrive at the relation

$$\log P = at + C \tag{7}$$

where C is an arbitrary constant and "log" denotes the natural logarithm; we may write log P rather than log $|P|$ because we know that population will always be nonnegative. If the population is known at some particular instant, then the value of C is easily computed. If, for example, $P = P_0$ at time $t = 0$, then $\log P_0 = aO + C = C$. Thus,

$$\log P = at + \log P_0 \tag{8}$$

Exponentiating each side of this equation yields an explicit relation between population P and time t:

$$P = P_0 e^{at} \tag{9}$$

The behavior of this function depends on the sign of a. If a is positive, then P is a steadily increasing, unbounded function of t. If a is zero, then P is the constant function whose value is P_0 for all t. If a is negative, then P is a steadily decreasing function of t that approaches zero as t grows large. See Fig. 3.1 for graphs of these three possibilities.

Exponential growth The constant a would be positive in the case of a pure birth process or if the birth rate b exceeded the death rate d. The model then predicts that the population will steadily increase and become indefinitely large. Ecologists would say that the population is undergoing *exponential growth*. Since the model asserts that there is no limit to the number of individuals in this population, it is clear that the model is not a completely realistic picture.

Before this model is scrapped, however, let us note that it may be a realistic one for the growth of some populations over relatively short time intervals. As an example of this, the population of the United States during the period from 1790 to 1860 grew at just such an exponential pace. See Fig. 3.2. To get an idea of what the growth rate a was during this

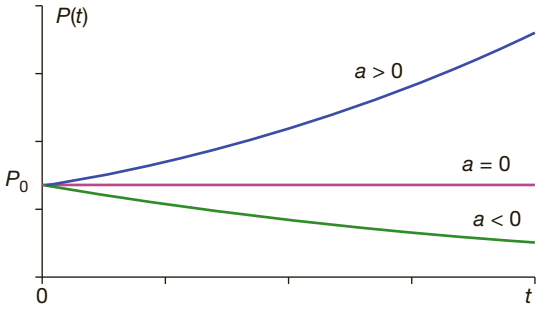


FIGURE 3.1 The graph $P = P_0 e^{at}$. The shape of the curve depends on the sign of a.

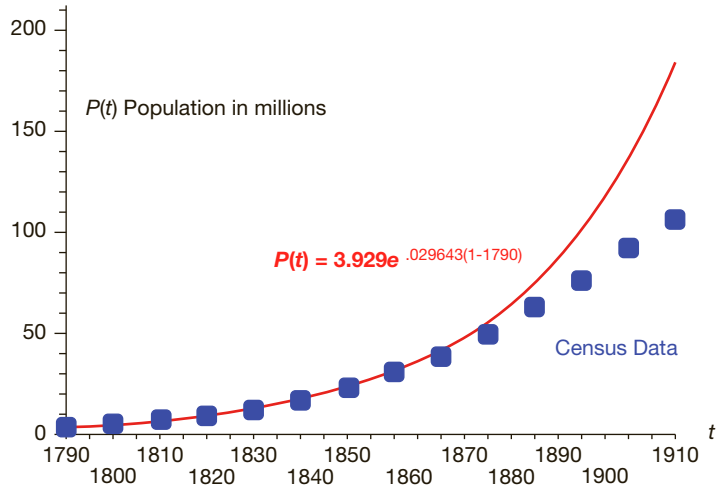


FIGURE 3.2 A comparison of population growth in the United States from 1790 to 1920 with the exponential curve $P(t) = 3.929e^{.029643(1790-t)}$. The "fit" is extremely close for the period 1790–1860.

period, let $t = 0$ correspond to 1790 and $P_0 = 3.929$ million, the population counted in the 1790 census. If the population growth actually followed an exponential curve, then the number of people in the United States in the year 1830 $(t = 40)$, for example, would satisfy

$$P_{40} = P_0 e^{40a} \tag{10}$$

This equation may be solved for a by taking logarithms:

$$\log P_{40} = \log P_0 + 40a \tag{11}$$

or

$$a = \frac{\log P_{40} - \log P_0}{40} \tag{12}$$

The census data give $P_{40} = 12.8607$ million. The value of a would then be $a = 0.029643$ and the equation for U.S. population growth would be

$$P(t) = 3.929e^{.029643(t - 1790)} \tag{13}$$

By the choice of constants, this model exactly predicts the population levels in the years 1790 and 1830. To test how well the model works as a predictor in other years, examine Table 3.1. Observe from this table that the predicted values of population are very close to the observed ones for the years 1790 to 1860. The largest error is less than 2 percent of the population. Since the census data, especially in the early years of the republic, was itself subject to many errors, this is as good a "fit" as we might reasonably expect.

As a long-term model of U.S. population growth, the model is not a very good one, as the data from 1870 through 2010 displayed in Table 3.1 show. Important factors such as wars, immigration, variations in the birth and death rates, and changes in the age structure of the population are missing from the model. Nevertheless, simple exponential growth is an accurate way of portraying the change in population in the United States during the early and middle nineteenth century. We may also conclude that if the pattern of growth established during that period continued until the present day, the population of the United States today would be well over 2.5 billion!

Scientists have obtained similar conclusions about exponential growth models for other living organisms. The simple model $dP/dt = aP$ is often very accurate when the environmental conditions are close to ideal: no natural enemies of the species are present, resources are unlimited, and there is sufficient space for the organisms to develop without interfering with each other.

It's not unusual for an exponential growth model to predict fairly accurately the population of a large nation over a several decade period. Looking again at U.S. census figures, if we take the years 1940 and 1950 as our initial data values, then the value of a for the exponential model is .013539. Table 3.2 displays a comparison between the predicted and observed population figures for the half-century after 1950.

Replacing the derivative with the difference in successive time periods transforms the exponential growth model $dP/dt = aP$ into the difference equation,

Table 3.1 A comparison between actual population based on census data in the United States and that predicted by exponential growth with rate 0.029643. The "error" term is found by subtracting the actual population from the predicted one. The "relative error" (not shown) is the error divided by the actual population and the "percent error" is the relative error multiplied by 100%.

Year	Predicted Population (in millions)	Observed Population (in millions)	Error (in millions)	Percent Error
1790	3.92921	3.92921	0.00	0.000
1800	5.28498	5.30848	−0.02	−0.443
1810	7.10861	7.23988	−0.13	−1.814
1820	9.56146	9.63845	−0.08	−0.800
1830	12.8607	12.8607	0.00	−0.002
1840	17.2983	17.0634	0.23	1.375
1850	23.2671	23.1919	0.07	0.322
1860	31.2956	31.4433	−0.15	−0.473
1870	42.0940	38.5584	3.53	9.167
1880	56.6191	50.1892	6.43	12.807
1890	76.1556	62.9798	13.17	20.916
1900	102.433	76.2122	26.22	34.400
1910	137.779	92.2285	45.54	49.381
1920	185.319	106.022	79.29	74.784
1930	249.266	123.203	126.05	102.309
1940	335.274	132.165	203.09	153.663
1950	450.959	151.326	299.61	197.988
1960	606.568	179.323	427.20	238.231
1970	815.865	203.302	612.51	301.279
1980	1097.39	226.542	870.76	384.370
1990	1476.04	248.710	1227.21	493.431
2000	1985.35	281.422	1703.76	605.413
2010	2670.03	308.746	2361.43	764.844

$$P_{i+1} - P_i = aP_i \tag{14}$$

which we may rewrite as

$$P_{i+1} = (1 + a)P_i$$

whose solution we found in Chapter 1 to be

$$P_k = (1 + a)^k P_0 \tag{16}$$

Let's use the discrete model to investigate how small changes in the initial population affect long-term predictions for the population. Small errors or "perturbations" in a constant

Table 3.2

Year	Predicted Population (in millions)	Observed Population (in millions)	Error	Percent Error
1940	132.165	132.165	0.000	0.000
1950	151.326	151.326	0.000	0.000
1960	173.265	179.323	–6.058	–3.378
1970	198.385	203.302	–4.917	–2.419
1980	227.146	226.542	0.604	0.267
1990	260.077	248.710	11.367	4.570
2000	297.782	281.422	16.360	5.813
2010	340.954	308.746	32.208	10.432

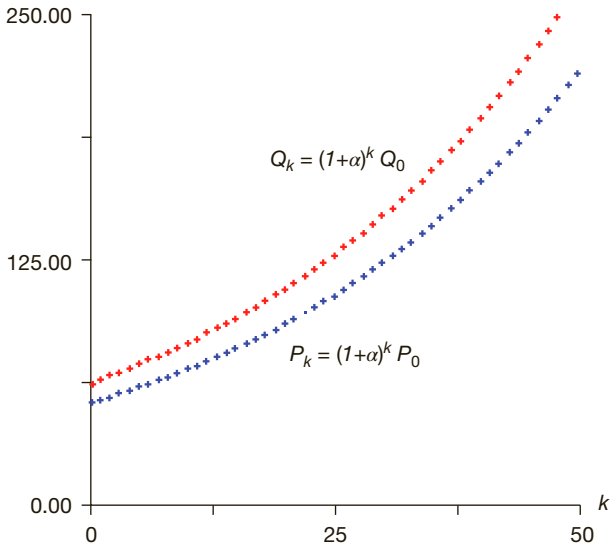


FIGURE 3.3 Dependence of the discrete exponential model on initial population. Here $a = .03$, $P_0 = 50$ and $Q_0 = 55$.

appearing in a mathematical model are inevitable. The constants represent real-world quantities whose measurement is necessarily inexact. Scientists are interested in the "robustness" of a model: are its long-term qualitative predictions essentially the same over a reasonable range of values for the constant parameters?

For the exponential model we have been discussing, a change in initial population from P_0 to Q_0 results in a prediction that the population Q_k after k time periods would be $Q_k = (1 + a)^k Q_0$. The difference between the two estimates $P_k - Q_k = (1 + a)^k (P_0 - Q_0)$ will grow large as k increases (assuming k is positive), but the relative difference,

$$\frac{P_k - Q_k}{P_k} = \frac{(1 + a)^k (P_0 - Q_0)}{(1 + a)^k P_0} = \frac{P_0 - Q_0}{P_0}$$

remains fixed. The graphs of P_k and Q_k will be qualitatively the same. Fig. 3.3 illustrates this idea where the gap between P_0 and Q_0 is even relatively large. The graphs are essentially parallel.

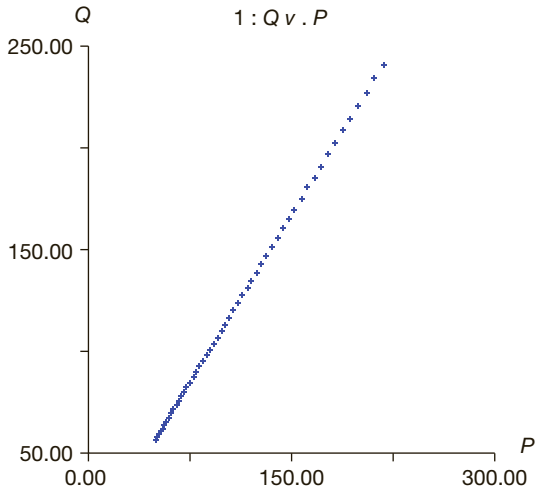


FIGURE 3.4 The points (P_k,Q_k) for the discrete exponential models with $a = .03$, $P_0 = 50$ and $Q_0 = 55$.

In Fig. 3.4 we plot the points (P_k,Q_k) for the discrete exponential model where $a = .03$, $P_0 = 50$, and $Q_0 = 55$. We see that the points do fall along the straight line $Q = \frac{Q_0}{P_0}P$.

III. Exponential Decay

If the sign of a is negative, then the function $P(t) = P_0e^{at}$ remains positive for all values of t but steadily decreases toward zero as t increases. If the population under consideration consists of all individuals belonging to a particular species, then the model predicts that the species will become extinct as all the individuals will eventually die off. As noted above, the constant a would be negative if the death rate exceeds the birth rate or if the assumptions of the pure death process are valid.

Unfortunately, there are conditions in our environment today that make a pure death process quite likely for the future growth of some species. The extensive use of pesticides, particularly DDT, in the mid-twentieth century had the unexpected consequence of drastically reducing the live birth rate of certain species of birds. One of the best-documented studies concerns the plight of the peregrine falcon, a bird of prey that once bred on cliff sides across the United States. The extensive use of DDT began in 1946, and the first signs of decrease in the peregrine population were noted within a year. In the 23 breeding seasons between 1947 and 1970—during which time DDT and similar persistent pesticides were abundantly used—the peregrine had become all but extinct as a breeding bird in the continental United States. Research showed that DDT absorbed by the birds inhibited an enzyme that facilitates the transporting of calcium from the blood to the site of eggshell production in the oviduct. As a result, the falcons lay thinner eggs, which crack under their weight when they brood them. Since the number of live births dropped dramatically while the death rate among adults remained essentially unchanged, it is not surprising that in some areas of the country less than 10 percent of the pre-pesticide breeding population remains. Eventually, the use of DDT was diminished and banned for some uses in 1972; residual DDT in the environment today, however, continues to contaminate peregrine falcons. Fortunately, peregrine falcons have steadily increased in number and consequently are no longer on the Endangered Species List.

The decay of radioactive material is another example of pure exponential decay, since the number of atoms that decompose in a given unit of time is proportional to the total number present. The rate of decay of a radioactive element is often expressed in terms of its *half-life*, the time required for a quantity of the element to decrease by a factor of one-half. In terms of the function $P(t) = P_0 e^{at}$, this number is given by $(-\log 2)/a$, which is independent of P_0.

In the late 1940s, Willard F. Libby discovered radiocarbon, a radioactive isotope of carbon with a half-life of approximately 5,600 years. The ratio of radioactive to nonradioactive carbon present in all living organisms has remained essentially constant over many centuries. When the organism dies, it stops absorbing new radiocarbon, so that the ratio decreases exponentially over the years. If an old bit of charcoal has half the radioactivity of a living tree, then it came from a tree that died about 5,600 years ago.

Libby and his coworkers developed the technique of radiocarbon dating to determine the ages of many objects dating back as much as 50,000 years. This technique has been of great significance to archeologists and anthropologists, whose use of radiocarbon dating and other observations indicate, for example, that humans arrived in the Western Hemisphere only about 11,500 years ago.

The common isotope of uranium has a half-life of 4.5 billion years, while rubidium decays into strontium with a half-life of 50 billion years. Using a dating technique based on the exponential decay of these radioactive elements, geologists have determined the ages of rocks found on the earth and on the surface of the moon. From these, they are obtaining a better picture of the development of our planet.

IV. Logistic Population Growth

A. The Logistic Model

The basic assumption of the pure exponential model is that the rate of increase of population is proportional to the size of the population—that is, the rate is a constant, independent of the size of the population. The model assumes that sufficient resources are available to sustain any level of population so that there is no interference between individuals in the population. These assumptions are not very realistic. Every species of organism inhabits some restricted environment, with a finite amount of space and a limited supply of resources. The environment has a *carrying capacity*, an upper limit on the number of individual organisms that can exist on the available resources. As the size of the population gets closer to this carrying capacity, its rate of growth must slow down. Any realistic model of population dynamics should reflect this feature. This section examines a mathematical model that attempts to do this.

Briefly stated, the argument in the paragraph above is that the rate of growth is not constant, but rather is dependent on the size of the population. The mathematical model should then assert that the rate of population is in fact a function of the population; mathematically, the statement looks like

$$\frac{dP}{dt} = f(P) \tag{17}$$

where f is some function of population size P. How should f be selected? If the population ever reaches a zero level, then of course it will always remain at zero. Hence, the function f

should have the property that $f(0) = 0$. Suppose we write the function f as $f(P) = Pg(P)$ where g is also a function of P. Then $f(0) = 0g(0) = 0$ regardless of the form of the function g. Note that we can think of g as the per capita growth rate, $g(P) = P'/P$. How then should g be selected?

The idea that rate of growth will slow down as population gets larger and larger can be captured by the condition that $g'(P)$ be negative. The simplest model is then obtained by making the function g as simple as possible—namely, assume that g is a linear function,

$$g(P) = a - bP \tag{18}$$

where a and b are positive constants. Then the model assumes the form

$$\frac{dP}{dt} = P(a - bP) \tag{19}$$

This assertion is called the *logistic equation* or the *Verhulst-Pearl equation*.

Note that this derivation of the logistic model does not make explicit use of the carrying capacity of the environment. There are several other ways of arriving at this model. We will outline one path: suppose that aP is the rate at which the population would increase if the environment possessed unlimited space and resources. Then we might assume that the actual growth rate is the potential rate multiplied by a factor measuring the proportion of the maximum attainable size of the population that is still unrealized. If M is the maximum possible population size in the environment, then $M - P$ is the amount of growth still available, and $(M - P)/M$ would be the fraction of maximum attainable size still possible. The assumption is then that the actual rate of growth is $aP(M - P)/M$. The differential equation expressing the model would be

$$\frac{dP}{dt} = P\frac{M - P}{M} = aP - \frac{a}{M}P^2 \tag{20}$$

which is easily recognized as the Verhulst-Pearl equation.

Before we begin a careful mathematical analysis of the logistic equation, it is useful to examine its direction fields. Fig. 3.5 shows the direction fields for the particular logistic equation $\frac{dP}{dt} = P(a - bP)$, where $a = .05$ and $b = .00005$. Note that $a/b = 1000$. We see that the population P appears to approach 1,000 in the long term, whether we begin with P above or below 1,000. The initial change in P is large, as the arrows near $t = 0$ are nearly vertical. As time progresses, the tangent lines begin to flatten out and become nearly horizontal.

B. Mathematical Analysis

The logistic equation may be solved by a nice application of the technique of partial fraction decomposition. The differential equation

$$\frac{dP}{dt} = P(a - bP)$$

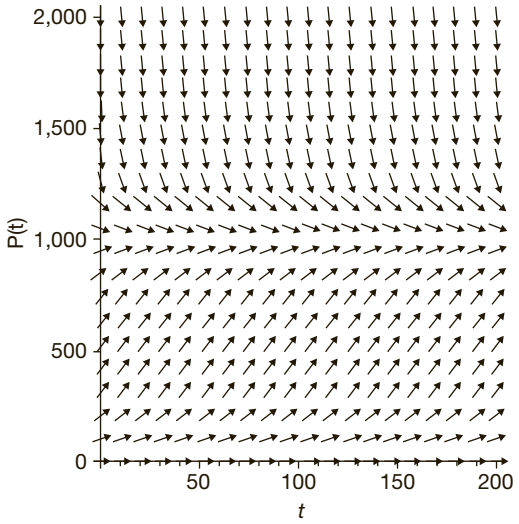


FIGURE 3.5 Direction field for a logistic equation.

may be written in the equivalent form

$$\int \frac{1}{P(a-bP)}\,dP = \int 1\,dt \tag{21}$$

after separating variables and integrating.

Now the fraction $\frac{1}{P(a-bP)}$ may be decomposed as

$$\frac{1}{P(a-bP)} = \frac{\frac{1}{a}}{P} + \frac{\frac{b}{a}}{a-bP} \tag{22}$$

so that we have the equivalent integration problem

$$\int \frac{1}{P} + \frac{b}{a-bP}\,dP = \int a\,dt \tag{23}$$

Simple integration then yields

$$\log P - \log(a-bP) = at + C \tag{24}$$

where C is a constant of integration. We may rewrite this last equation as

$$\log \frac{P}{a-bP} = at + C \tag{25}$$

Exponentiation of each side gives

$$\frac{P}{a-bP} = Ke^{at} \tag{26}$$

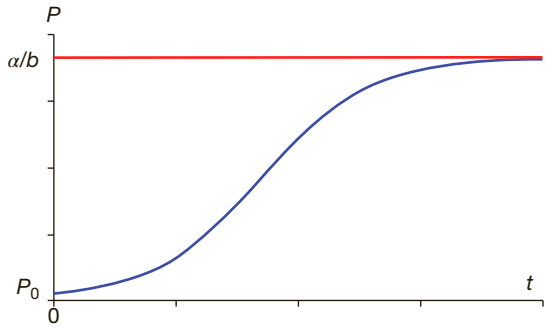


FIGURE 3.6 The logistic curve $P = k/\left(1 + e^{(d-at)}\right)$.

where K is the constant e^C. It is useful to rewrite this equation as

$$P = \frac{aKe^{at}}{1 + bKe^{at}} = \frac{a}{b + (1+K)e^{-at}} = \frac{a/b}{1 + (1/bK)e^{-at}} = \frac{k}{1 + e^{d-at}} \tag{27}$$

where $d = -\log(bK)$ and $k = a/b$. Since a is positive, e^{-at} tends to 0 as t increases. Thus, $\lim_{t\to\infty} P(t) = k = a/b$. The logistic model then predicts that population will increase and asymptotically approach the carrying capacity a/b. Note that at the capacity level of $P = a/b$, the logistic equation gives $dP/dt = 0$. See Fig. 3.6 for a graph of the population as a function of time. This curve is called the *logistic curve*, and resembles an elongated letter S.

Another useful way to examine logistic growth comes if we make the change of variable $Q = \frac{b}{a}P$ in the differential equation $\frac{dP}{dt} = P(a-b)P$. We can interpret $Q(t)$ as the fraction of the carrying capacity the population has reached at time t, since $Q = 0$ when $P = 0$ and $Q = 1$ when $P = \frac{a}{b}$, the carrying capacity. Then we have

$$\frac{dQ}{dt} = \frac{b}{a}\frac{dP}{dt} = \frac{b}{a}P(a - bP) = Q(a - aQ) = aQ(1 - Q).$$

or, more simply,

$$\frac{dQ}{dt} = aQ(1 - Q) \tag{28}$$

It is straightforward to show that $Q(t) = \frac{1}{1 + Ce^{-at}}$ where the constant C is equal to $\frac{1}{Q(0)} - 1$.

C. Testing the Logistic Model

Laboratory experiments with a variety of species have shown that the growth of many populations, under appropriate conditions, follows the logistic curve. As a second model of growth of the U.S. population, consider the logistic equation. We shall show that this model gives an accurate portrait of the changes in the nation's population for much of its history.

In each of its forms, the logistic model of population as an explicit function of time contains three constants. To test the equation as a model of population growth in the United States, we must assign numerical values to these constants. It is possible to do this if the populations P_0, P_1, and P_2 are known at the three times t_0, t_1, and t_2. To this end, rewrite the

equation of the logistic curve as

$$d - at = \log\left(\frac{k-P}{P}\right) \tag{29}$$

where a, d, and k are the constants to be determined. With the equation in this form, we have

$$d - at_0 = \log\left(\frac{k-P_0}{P_0}\right) = A \tag{30}$$

$$d - at_1 = \log\left(\frac{k-P_1}{P_1}\right) = B \tag{31}$$

and

$$d - at_2 = \log\left(\frac{k-P_2}{P_2}\right) = C \tag{32}$$

These equations yield the relationships

$$\begin{aligned} B - A &= a(t_0 - t_1) \qquad & \mathrm{C} - A &= a(t_0 - t_2) \\ a &= \frac{B-A}{t_0 - t_1} \qquad \text{and} & d &= at_0 + A. \end{aligned} \tag{33}$$

From the first pair of equations, we have

$$(t_0 - t_2)(B - A) = (t_0 - t_1)(C - A) \tag{34}$$

For simplicity, suppose we choose equally spaced dates so that $t_2 - t_1 = t_1 - t_0$. Then $2(B - A) = C - A$ or $2B = A + C$. This equation gives

$$\log\left(\frac{k-P_1}{P_1}\right)^2 = \log\left(\frac{k-P_0}{P_0}\right) + \log\left(\frac{k-P_2}{P_2}\right) \tag{35}$$

or

$$\left(\frac{k-P_1}{P_1}\right)^2 = \left(\frac{k-P_0}{P_0}\right)\left(\frac{k-P_2}{P_2}\right) \tag{36}$$

Since P_0, P_1, and P_2 are known, we have a quadratic equation in k. This has two roots, $k = 0$ and

$$k = \frac{P_1(2P_0P_2 - P_0P_1 - P_1P_2)}{P_0P_2 - P_1^2} \tag{37}$$

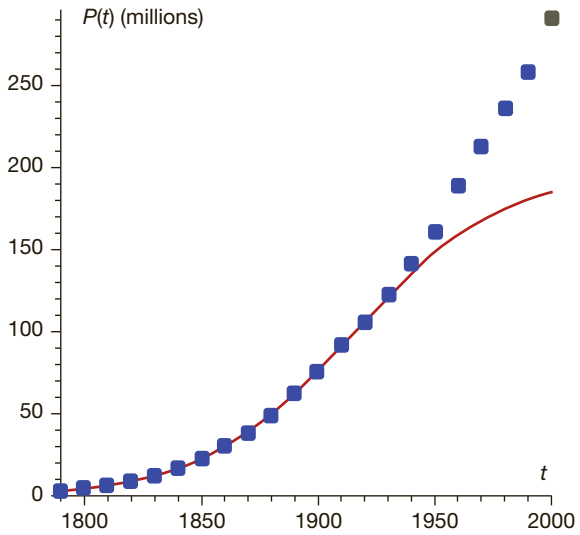


FIGURE 3.7 A comparison of population growth in the United States from 1790 to 2000 and the logistic curve $P(t) = 197.274/\left(1 + e^{(59.97-.031t)}\right)$. As in Fig. 3.3, the heavy dots represent actual population levels.

The root $k = 0$ corresponds to the fact that if the population ever reaches zero, it will remain there forever. The nonzero root for k gives the carrying capacity. Once k is computed from the known values P_0, P_1, and P_2, then A, B, and C are easily calculated. From these the values of a and d can then be found. As an example, suppose the population figures of the censuses in 1790, 1850, and 1910 are used to determine the constants. The value of k turns out to be 198.947 and the predicted equation for population growth in the United States looks like

$$P(t) = \frac{198.947}{1 + e^{(59.9722 - 0.0313227t)}} \text{millions} \tag{38}$$

Comparisons between the predictions of this model and the actual population figures are given in Fig. 3.7 and in Table 3.3. The table shows that the model gives an excellent portrayal of the changes in U.S. population from 1790 through 1950, the largest deviation being less than 4%. The model fails, however, after the middle of the 20th century. It does not predict the increase in the birth rate that led to the unexpected and unprecedented increase of 30 million Americans between 1950 and 1960. The model clearly has failed to include some factors that critically affected population changes in the last 60 years.

The dates chosen above that were used to determine the constants in the logistic equation were the ones selected by Raymond Pearl and Lowell J. Reed in a 1924 study of the U.S. population growth curve. Impressed by the closeness of the fit of the logistic equation to the census data from 1790 through 1920 (the only numbers available to them), they wrote [Pearl and Reed, 1920], "so far as we may rely upon present numerical values, the United States has already passed its period of most rapid population growth, unless there comes into play some factor not now known and which has never operated during the past history of the country to make the rate of growth more rapid. This latter contingency is improbable."

Table 3.3 A comparison between actual population in the United States and that predicted by a logistic equation. The "error" term is found by subtracting the actual population from the predicted one. The "percent error" is the error divided by the actual population.

Year	Predicted Population (in millions)	Observed Population (in millions)	Error	Percent Error
1790	3.929	3.929	0.000	0.01
1800	5.336	5.308	0.028	0.53
1810	7.228	7.240	−0.012	−0.17
1820	9.756	9.638	0.117	1.22
1830	13.108	12.861	0.247	1.92
1840	17.506	17.063	0.443	2.59
1850	23.194	23.192	0.002	0.01
1860	30.420	31.443	−1.023	−3.25
1870	39.393	38.558	0.834	2.16
1880	50.228	50.189	0.038	0.08
1890	62.864	62.980	−0.116	−0.18
1900	77.032	76.212	0.820	1.08
1910	92.235	92.228	0.006	0.01
1920	107.780	106.022	1.758	1.66
1930	122.927	123.203	−0.276	−0.22
1940	137.005	142.165	−5.160	−3.63
1950	149.526	161.326	−11.800	−7.31
1960	160.229	189.323	−29.094	−15.37
1970	169.078	213.302	−44.224	−20.73
1980	176.190	236.542	−60.352	−25.51
1990	181.783	258.710	−76.927	−29.73
2000	186.100	291.422	−105.322	−36.14

Pearl and Reed's estimate of an eventual population in the nation of slightly under 200 million was wrong, but not because they made a poor choice of sample years. Almost any triple of years selected for t_0, t_1, and t_2 that roughly coincide with early, middle, and contemporary dates for them would yield a particular form of the logistic curve with a similar property: the jump in population from 1950 to 1960 just does not parallel the climb of the logistic curve.

Suppose we focus on population growth in the United States during the 20th century only, ignoring what happened before 1900. If we fit a logistic model using the census data for the years 1930, 1960, and 1990, then the logistic curve is described by the equation

$$P(t) = \frac{640.26}{1 + e^{(32.9735 - 0.016342t)}} \tag{39}$$

Table 3.4

Year	Predicted Population	Observed Population	Error	Percent Error
1910	93.892	92.229	1.663	1.80
1920	107.755	106.022	1.733	1.63
1930	123.203	123.203	0.000	0.00
1940	140.283	132.165	8.118	6.14
1950	159.002	151.326	7.676	5.07
1960	179.323	179.323	0.000	0.00
1970	201.156	203.302	−2.146	−1.06
1980	224.353	226.542	−2.189	−0.97
1990	248.710	248.710	0.000	0.00
2000	273.969	281.422	−7.453	−2.65

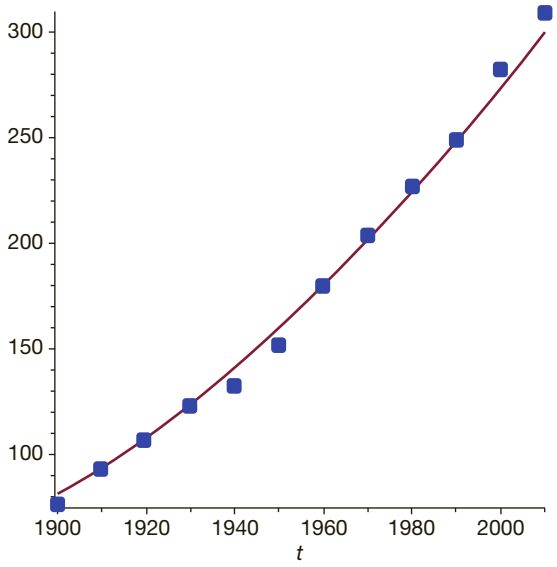


FIGURE 3.8 A comparison of population growth in the United States during the 20th century and the logistic curve. As in Fig. 3.6, the heavy dots represent observed population levels.

Table 3.4 and Fig. 3.8 show how closely this equation predicted the census figures for the years 1910−2010. Observe that the largest error is about 6%. Note also that the equation predicts that the U.S. population will approach a limiting value of 640.26 million as time advances. It will be interesting to see how accurate this prediction turns out to be. Incidentally, this model predicts a 2020 population of 325.959 million; the current U.S. Census Bureau projects a figure of 322.742 million.

A more accurate model of the growth of U.S. population can be obtained by refining the model in several different ways. The function $g(P)$—chosen to be linear in the logistic model—might be taken to be a polynomial of higher degree so that higher-order effects of the size of population on the growth rate could be included. Additional factors might be attached to the differential equation to incorporate the concept that the rate of change of population is not only a function of population but of time as well. We might also try to

include such factors as immigration, the medical and public health discoveries that have increased life expectancy by 15 years in the last half-century, changes in the age structure of the population, and the effects of depressions and periods of economic prosperity. The population might be divided into ethnically or religiously determined groups that display different birth rate patterns. Some of these approaches are outlined in the exercises.

Demographers are using increasingly complex and sophisticated mathematical models of both deterministic and probabilistic character to study changes in population growth in the past and to make projections about the future. An elementary probabilistic model of population growth is presented in Chapter 10.

V. The Discrete Model of Logistic Growth and Chaos

At the conclusion of Part B of Section IV, we examined the continuous model of logistic growth using Q as the variable that is the fraction of the carrying capacity:

$$\frac{dQ}{dt} = aQ(1 - Q) \tag{28}$$

A discrete version of this model would have the population in time period i satisfy the difference equation

$$Q_{i+1} - Q_i = aQ_i(1 - Q_i) \tag{40}$$

The behavior of the discrete logistic growth model turns out, surprisingly, to be quite dependent on the value of a.

For example, with $a = 1.7$ and $Q_0 = .3$, Fig. 3.9 shows Q_k rapidly approaching 1, as we would expect.

If a is increased to 2.2, however, then a very different behavior is observed. The values of Q_k appear to oscillate between two different values. For $Q_0 = 0.3$, these values are about 1.16 and 0.75. Fig. 3.10 displays the results.

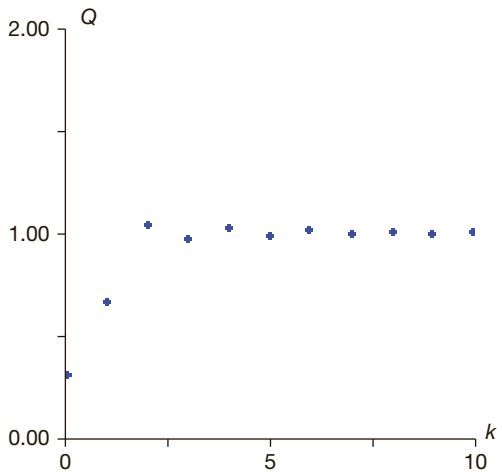


FIGURE 3.9 Behavior of the discrete logistic model with $a = 1.7$.

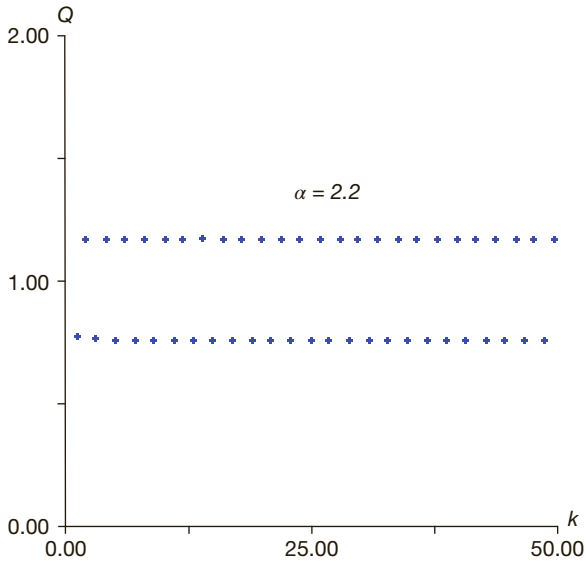


FIGURE 3.10 Behavior of the discrete logistic model with $a = 2.2$.

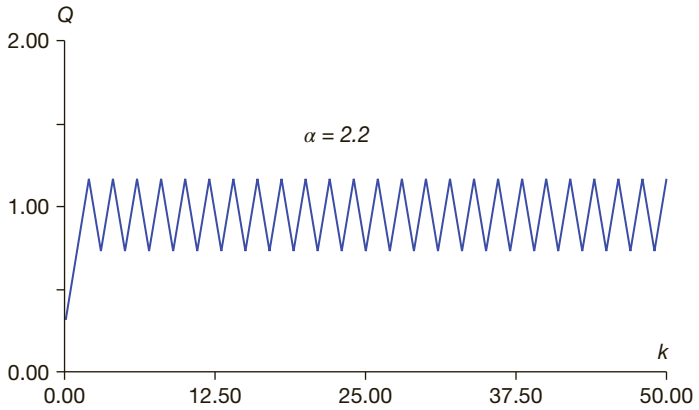


FIGURE 3.11 Connecting the dots for the discrete logistic models with $a = 2.2$.

Fig 3.11 shows the data if we connect the consecutive dots.

Table 3.5 displays some of the numerical values.

With a increased to 2.5, as k increases, the subsequent values of Q_k appear to rotate through values close to 1.22, 0.54, 1.16, and 0.70. Such a "4 cycle" is seen in Figs. 3.12 and 3.13; the numerical values are displayed in Table 3.6.

As a increases still more to a level of $a = 2.7$, even more bizarre patterns are seen in the values of Q_k. With $Q_0 = 0.5$, the first 200 values of Q_k are displayed in Fig. 3.14. They appear to be scattered almost at random.

If we connect points for consecutive values of k, we see that there may be a pattern underlying the distribution, although it is a highly irregular one. See Fig. 3.15.

Table 3.5

k	Q_k	k	Q_k	k	Q_k
0	0.30	10	1.16	20	1.16
1	0.76	11	0.75	21	0.75
2	1.16	12	1.16	22	1.16
3	0.75	13	0.75	23	0.75
4	1.16	14	1.16	24	1.16
5	0.75	15	0.75	25	0.75
6	1.16	16	1.16	26	1.16
7	0.75	17	0.75	27	0.75
8	1.16	18	1.16	28	1.16
9	0.75	19	0.75	29	0.75

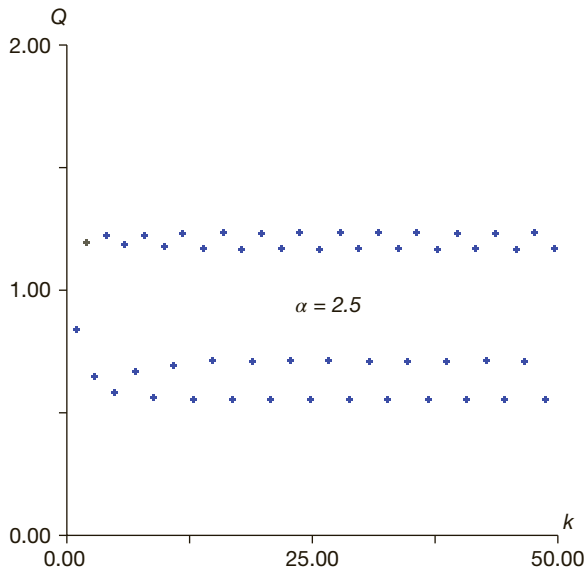


FIGURE 3.12

Table 3.7 displays the first 100 iterations of the discrete logistic model with $Q_0 = 0.3$ and $a = 2.7$. It is difficult to discern whether there is a cycle of values underlying the distribution.

We see then a major qualitative difference between the discrete exponential growth model $Q_{i+1} - Q_i = aQ_i$ and the discrete logistic growth model $Q_{i+1} - Q_i = aQ_i(1 - Q_i)$. The former model is called a *linear* dynamic system, because the difference between consecutive terms is a linear function of the ith term. In the logistic model, the difference is a quadratic function of the ith term. The logistic model is an example of a *nonlinear* dynamic system.

The discrete logistic model also displays an important condition known as *sensitive dependence on initial conditions*. In our examination of the linear exponential growth

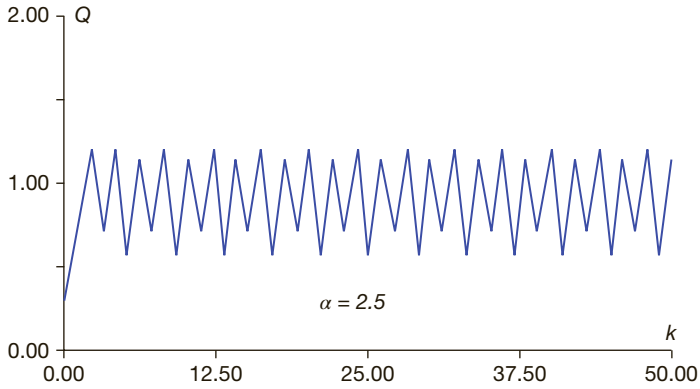


FIGURE 3.13

Table 3.6

k	Q_k	k	Q_k	k	Q_k
0	0.30	10	1.17	20	1.22
1	0.82	11	0.67	21	0.54
2	1.19	12	1.22	22	1.16
3	0.63	13	0.54	23	0.70
4	1.21	14	1.16	24	1.22
5	0.56	15	0.69	25	0.54
6	1.18	16	1.22	26	1.16
7	0.65	17	0.54	27	0.70
8	1.22	18	1.16	28	1.22
9	0.55	19	0.70	29	0.54

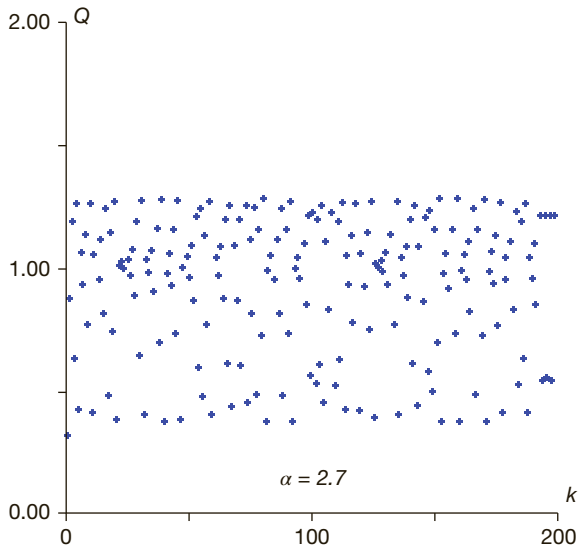


FIGURE 3.14

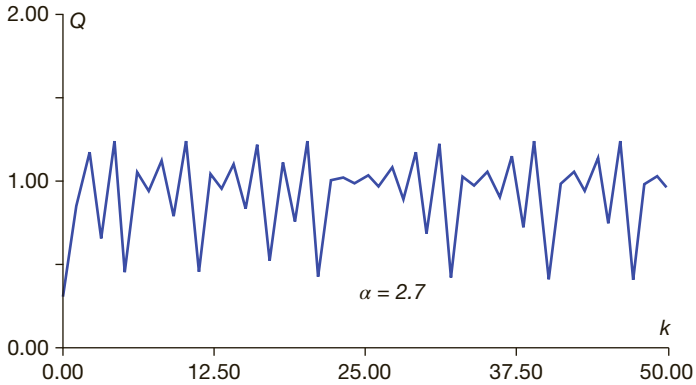


FIGURE 3.15

Table 3.7

k	Q_k	k	Q_k	k	Q_k	k	Q_k	k	Q_k
0	0.30	20	1.26	40	0.35	60	0.39	80	0.71
1	0.87	21	0.37	41	0.97	61	1.03	81	1.27
2	1.18	22	0.99	42	1.05	62	0.95	82	0.36
3	0.61	23	1.01	43	0.91	63	1.07	83	0.98
4	1.25	24	0.99	44	1.14	64	0.86	84	1.04
5	0.40	25	1.02	45	0.72	65	1.18	85	0.93
6	1.04	26	0.96	46	1.26	66	0.60	86	1.10
7	0.92	27	1.07	47	0.36	67	1.25	87	0.80
8	1.12	28	0.87	48	0.98	68	0.42	88	1.23
9	0.76	29	1.17	49	1.03	69	1.08	89	0.47
10	1.25	30	0.62	50	0.95	70	0.86	90	1.14
11	0.39	31	1.26	51	1.08	71	1.19	91	0.72
12	1.04	32	0.38	52	0.85	72	0.58	92	1.27
13	0.93	33	1.02	53	1.19	73	1.24	93	0.36
14	1.10	34	0.96	54	0.57	74	0.44	94	0.98
15	0.80	35	1.06	55	1.23	75	1.10	95	1.03
16	1.23	36	0.89	56	0.45	76	0.80	96	0.95
17	0.46	37	1.15	57	1.12	77	1.23	97	1.08
18	1.13	38	0.68	58	0.75	78	0.47	98	0.84
19	0.73	39	1.27	59	1.26	79	1.14	99	1.20

model, we saw that if the initial values P_0 and Q_0 were reasonably close together, this did not change the long-term qualitative nature of the curve of the subsequent values. The graphs were essentially parallel, and one could quite easily predict, for each k, the value of Q_k from the value of P_k. Thus, long-term behavior was relatively insensitive to perturbations in the initial conditions. The outcome for the discrete logistic model is very, very different.

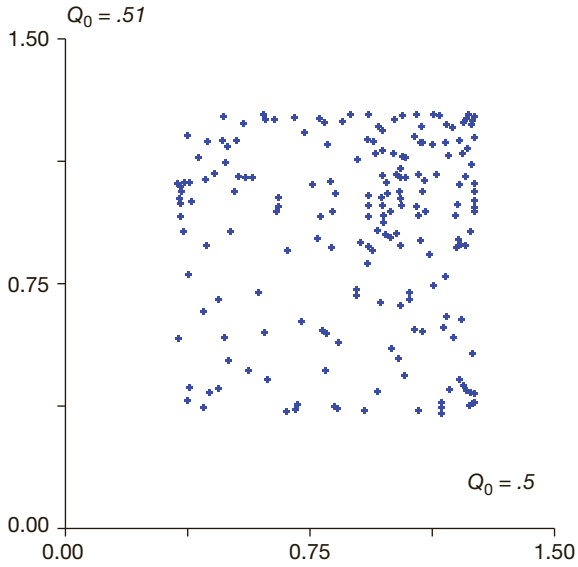


FIGURE 3.16

To illustrate the sensitive dependence on initial conditions, consider the logistic models

$$Q_{i+1} - Q_i = aQ_i(1 - Q_i) \quad \text{and} \quad P_{i+1} - P_i = aP_i(1 - P_i).$$

With $a = 2.7$ in both equations, while $Q_0 = 0.5$ and $P_0 = 0.51$. Fig. 3.16 shows a plot of the points (P_k, Q_k). Instead of these points falling along a straight line as they did for the exponential growth model (see Fig. 3.4) or along a simple one-dimensional curve, the points pepper the plane. Some clusters of nearby points appear, but there are many relatively isolated points.

If we plot the relative differences, $\frac{Q_k - P_k}{Q_k}$ versus k, we see that there is no consistency. Sometimes the relative difference is quite small, but other times it exceeds a factor of 2.

Table 3.8 provides a numerical picture of this phenomenon. After 50 steps, for example, Q_{50} is less than half the value of P_{50}, but Q_{130} and P_{130} are close together. By the time k is 200, P_{200} is 250% bigger than Q_{200}. Tiny changes in the initial conditions may result in very large changes in the values of the variables after a moderate number of steps.

Sensitive dependence on initial conditions is a characteristic of the mathematical concept known as *chaos theory*. Chaos theory deals both with showing how nonlinear deterministic systems can give rise to outcomes that appear erratic, random, and unpredictable and with discovering that what appears to be a chaotic output may actually contain complex patterns generated by relatively simple nonlinear discrete equations.

Many nonlinear systems, both discrete and continuous, which model important physical phenomena exhibit sensitive dependence on initial conditions. Henri Poincaré (1854–1912) first observed this more than a century ago when he attempted to analyze models for the three-body problem (see Chapter 1), but he lacked the computers necessary to investigate this property.

Table 3.8

k	Q_k	P_k	k	Q_k	P_k
0	0.50	0.51	110	0.86	1.25
10	0.56	0.48	120	0.81	1.17
20	1.14	1.27	130	1.10	1.18
30	0.48	0.70	140	0.90	1.13
40	1.27	1.00	150	1.22	1.19
50	0.61	1.26	160	1.13	1.17
60	0.79	0.60	170	0.94	1.27
70	0.53	1.03	180	0.84	1.03
80	1.02	1.15	190	0.62	1.25
90	1.16	0.36	200	0.36	0.91
100	0.46	1.09			

In the early 1960s, Edward Lorenz (1917−2008) observed sensitive dependence on initial conditions in a system of differential equations arising from the study of meteorological conditions. Starting off the identical system from two very slightly different initial values produced, after a relatively short time, overwhelming differences in the values of the basic variables. Since exact measurement of initial conditions is physically impossible, Lorenz concluded that accurate long-term weather prediction was an impossibility. In a paper in 1963 given to the New York Academy of Sciences Lorenz states:

> *One meteorologist remarked that if the theory were correct, one flap of a seagull's wings would be enough to alter the course of the weather forever.*

The sea gull evolved into the perhaps more poetic butterfly some time later, possibly because some of Lorenz solution curves suggested the shape of a butterfly. At the December 1972 meeting of the American Association for the Advancement of Science in Washington, D.C. Lorenz unveiled the new metaphor in the title of his talk: "Predictability: Does the Flap of a Butterfly's Wings in Brazil Set Off a Tornado in Texas?"

Today, chaos theory is a rapidly developing field that holds the promise of making significant contributions to our understanding of the physical and social sciences. Writing on the impact of chaos theory in physics, Trinh Xuan Thuan observed:

> *The last bastion of certainty collapsed at the end of the century: The emerging field of chaos eliminated once and for all the Newtonian and Laplacian tent of Nature's unconditional determinism. Before the advent of chaos, the operative word was* order. *The word* disorder *was anathema and banned from the language of science. Anything apt to exhibit irregularity or disorder was considered a monstrosity. The science of chaos changed all that. It introduced irregularity in regularity, disorder in order. It captured the imagination not only of scientists but also of the public at large, because chaos theory deals with objects on a human scale and speaks to everyday experiences.*

VI. The Allee Effect

A. The Allee Effect

One property of the logistic model is that the *per capita* growth rate is a strictly decreasing function of the population. The per capita growth rate is the rate of growth divided by the population size. For logistic growth, the per capita growth rate is

$$\frac{dP/dt}{P} = \frac{P(a-bP)}{P} = a - bP$$

Since a and b are positive constants, the graph of the per capita growth rate is a straight line of negative slope. The logistic model asserts that no matter what the population size is, increasing the number of individuals will always lower per capita growth. This property derives from the assumption that as population increases, there is always more competition for limited resources.

In the real world, however, we often observe more complicated behavior. When population density is very small, individuals may have more difficulty finding mates than they would if the population were larger. If a particular species hunts in packs, then when the population is very small, obtaining food may be more difficult resulting in poorer health, weakness, and inability to procreate. Growth rates may also be lower when there aren't enough individuals to form an effective group to defend against predators. For such animals, the per capita growth rate might actually increase as numbers build up from scarcity to a more abundant population.

The American zoologist and ecologist W. Clyde Allee (1885−1995) observed such behavior in a number of species. In his experiments, Allee noticed that goldfish grew more rapidly when there were more individuals in the tank, and that for certain land isopods, undercrowding, rather than competition, forced more limited population growth.

The "classical" view of population dynamics posited that because of competition for resources, a population will have a reduced overall growth rate at higher density and increased growth rate at lower density. Allee introduced the idea, now called the *Allee effect*, that the reverse holds true when the population density is low. In the extreme case, if population drops below a critical level, the species will become extinct.

There are several ways in which the Allee effect can be incorporated into a population growth model. Here is a simple extension of the logistic model:

$$\frac{dP}{dt} = P(a-bP)(P-c)$$

where the constant c is less than the carrying capacity a/b. It is useful to rewrite this equation as

$$\frac{dP}{dt} = P(a-bP)(P-c) = bP\left(\frac{a}{b} - P\right)(P-c) = bP(K-P)(P-c)$$

where K is the carrying capacity. If $0 < c < K$, then we can regard c as a *threshold*; if the population drops below c, then it will go extinct.

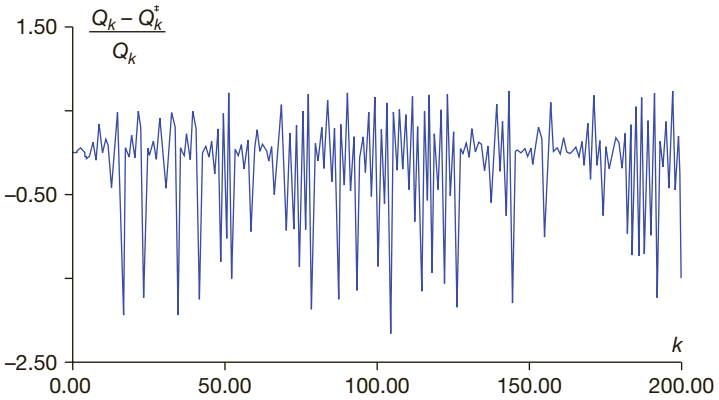


FIGURE 3.17

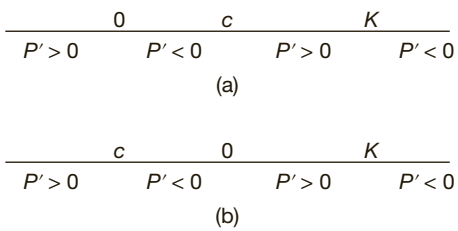


FIGURE 3.18

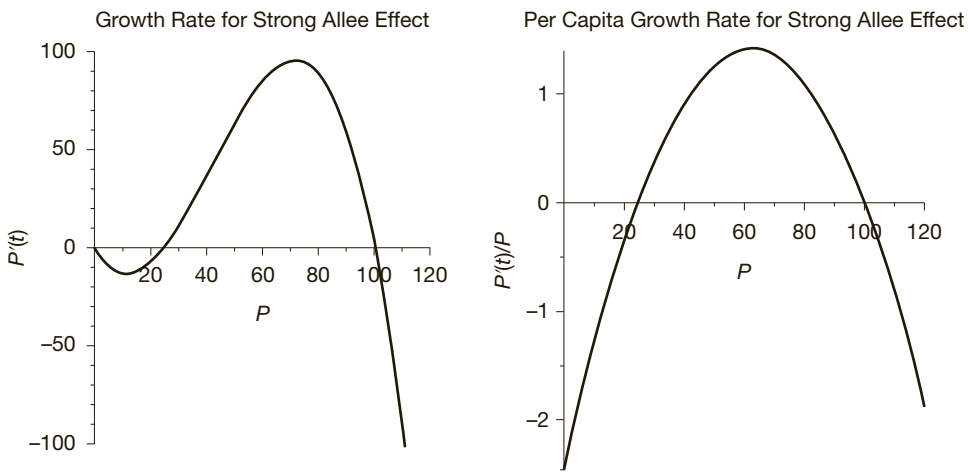


FIGURE 3.19 The growth rate and per capita growth rate for a logistic model with a strong Allee effect: $dP/dt = b(K - P)(P - c)$. Here $b = .001$, $K = 100$, $c = 25$.

We can gain more insight into the behavior of this logistic/Allee model by examining how the sign of $P' = dP/dt$ varies with P. Fig. 3.18 shows that there are two cases: $c > 0$ and $c < 0$.

When $c < 0$, we speak of a *weak Allee effect*, and when $c > 0$, it's called a *strong Allee effect*.

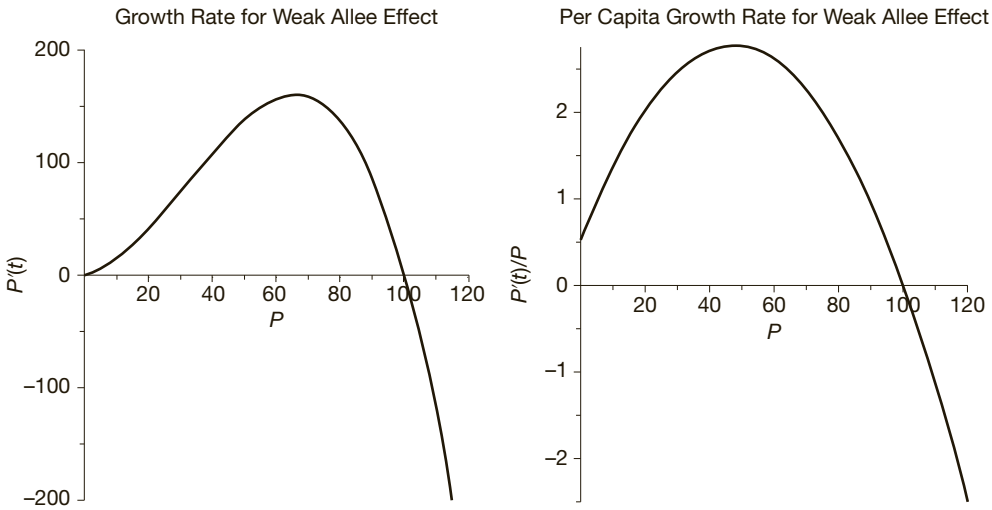


FIGURE 3.20 The growth rate and per capita growth rate for a logistic model with a weak Allee effect: $dP/dt = b(K - P)(P - c)$. Here $b = .001$, $K = 100$, $c = -5$.

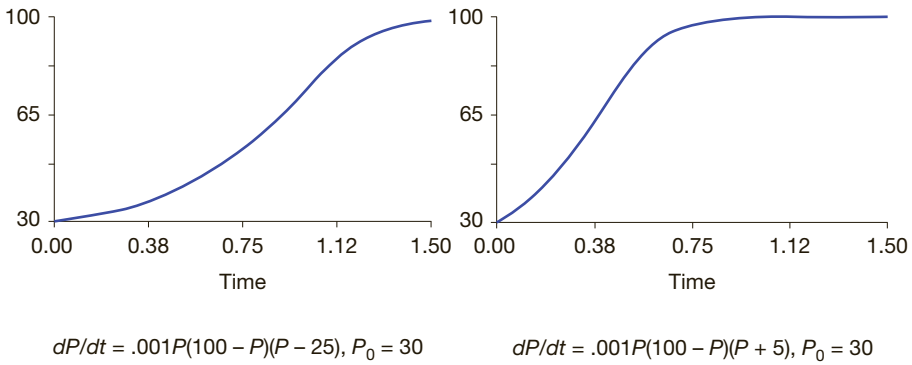


FIGURE 3.21 Trajectories of logistic model with Allee effects.

VII. Historical and Biographical Notes

A. Thomas Robert Malthus

"Explanations of population changes have been advanced by writers of many nationalities, religions, occupational specialties, and educational attainments," wrote Ralph Thomlinson in a [1965] study of population dynamics:

> *Independent and dependent variables which have been used for this purpose include total population, density, fertility, mortality, migration, climate, food, topography, energy sources, standard of living, level of aspiration, urbanization, degree of worldliness, transport facilities, technological development, balance of trade, genetic deterioration, age-sex distribution, socioeconomic class, religious belief, type of government, alcohol consumption, state of knowledge, and various combinations thereof. Most of these generalizations are over-simplified or obsolete; some are generally viewed as ludicrous; and a few are brilliant contributions to man's understanding of his own propagation, wandering, and demise.*

Prior to the 18th century, most of the writing on population was marred by superficial observations, strong doses of moral pronouncements, and a general failure to distinguish between folklore and factual evidence. It was believed by many, for example, that intellectual pursuits tended to diminish the power of procreation, that prostitutes could not conceive, and that "idiots bred like rabbits."

The central figure in the history of population theory is the Reverend Thomas Robert Malthus (1766–1834). Malthus, the second of eight children of an English country gentleman, won honors as a mathematics graduate of Cambridge University. He was ordained a minister in the Church of England in 1788. A year after his marriage in 1804, Malthus accepted an appointment as professor of history and political economy in East-Indian College, Hailebury, England. In addition to his teaching duties, Malthus published three important books on political economy, many pamphlets and tracts, and six editions of his famous essay on population.

Thomas Malthus. From a portrait by John Linnell.

Malthus's work on economics included a general theory of rent and the distribution of wealth, which has been cited as one of the foundation stones of modem economic thought. In biology, Charles Darwin wrote of his debt to Malthus for the phrase "struggle for survival" and for the concept that species may alter through selection. "In October 1838, I happened to read for amusement 'Malthus on Population'," wrote Darwin, "and being very well prepared to appreciate the struggle for existence which everywhere goes on, from long continued observation of the habits of animals and plants, it at once struck me that under these circumstances favorable variations would tend to be preserved, and unfavorable ones to be destroyed. The result would be the formation of a new species. Here then I had a theory by which to work."*

The first edition of his essay on population was published in 1798 under the title, "An Essay on the Principle of Population as It Affects the Future Improvement of Society, with Remarks on the Speculations of Mr. Godwin, M. Condorcet, and Other Writers." Condorcet

*_The Autobiography of Charles Darwin_, New York: Harcourt, Brace, 1959, p. 120.

and Godwin had each published works in 1793 emphasizing their optimistic beliefs in the perfectibility of man and society. They foresaw a day when inequality would be eliminated along with crime, disease, and war, a period in which reason would hold sway over emotion and base instincts.

Malthus had a contrary view. Social progress was illusory: "The structure of society, in its great features, will probably always remain unchanged." He thought man to be a lazy creature by nature, impelled to productive work only by a wife and children who needed food and shelter.

The essay on population sought to develop the consequences of two fundamental observations:

> *First, that food is necessary to the existence of man. Secondly, that the passion between the sexes is necessary and will remain in its present state. These two laws ever since we have had any knowledge of mankind appear to have been fixed laws of our nature; and as we have not hitherto seen any alteration in them, we have no right to conclude that they will ever cease to be what they now are.*

In 1826, a sixth edition of the essay appeared. It was titled "An Essay on the Principle of Population," or "A View of Its Past and Present Effects on Human Happiness, with an Inquiry into Our Prospects Respecting the Future Removal or Mitigation of the Evils Which It Occasions." Here Malthus summarizes three important conclusions:

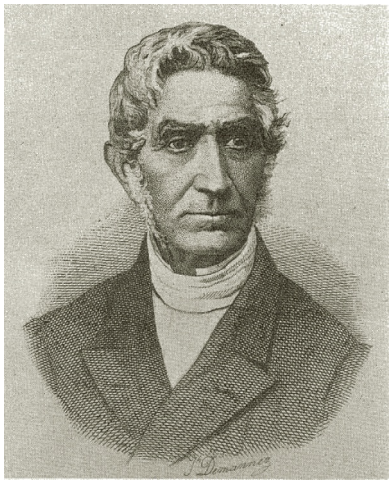
1. Population is necessarily limited by the means of subsistence.
2. Population invariably increases where the means of subsistence increase, unless prevented by some very powerful and obvious checks.
3. These checks, and the checks that repress the superior power of population, and keep its effects on a level with the means of subsistence, are all resolvable into moral restraint, vice, and misery.

During his lifetime and in subsequent generations, Malthus's essay provoked considerable controversy and debate. In tracing the development of a mathematical modeling approach to population, the next important contributor was one of the participants in this debate, Adolphe Quetelet.

B. Lambert Adolphe Jacques Quetelet

Mathematician, astronomer, sociologist, poet, statistician, physicist, man of letters, meteorologist—it is difficult to fit Quetelet into a single category. "Nature had endowed him not only with a vivid imagination and a mind of power, but also with the precious gift of indomitable perseverance," wrote Edouard Mailly [1875, 169].

Quetelet was born in Ghent, Belgium, on February 22, 1796, and educated in the local schools. In 1819 he received the first Doctor of Science degree awarded by the University of Ghent and shortly thereafter he assumed a professorship of mathematics at the Brussels Athenaeum. In Brussels, he quickly established many associations with the artists and writers of the area, became a member of the reading committee for the royal theater, and published many poems in the annual almanac of the local literary society.



Lambert A. J. Quetelet. From a portrait by J. Demannez. Copyright Bibliothèque royale Albert Ier, Brussels (Cabinet des Estampes).

His academic lectures, whether on elementary mathematics, calculus, experimental physics, or astronomy, were well received. "He was very highly esteemed by his pupils," Mailly [1875, 172] noted. "There was something about him at once imposing and amiable, while there was a complete absence of anything like pedantry or haughtiness. Although marked with smallpox, his physiognomy was refined and impressive; it was only necessary to fix his large dark eyes, surmounted with heavy black brows, upon the refractory, to insure at once silence and submission."

Although much of his early research was devoted to questions of primarily mathematical interest, Quetelet soon turned to applications. He superintended the construction of the Royal Astronomical Observatory and served as its first director from 1828 until his death in 1874, 5 days short of his seventy-eighth birthday. Work at the observatory included cataloging of stars, a careful study of atmospheric waves for the purposes of improving meteorology, and measurements of terrestrial magnetism.

Quetelet's contributions to the development of the social sciences derive from his efforts to apply probability and statistics to the study of man. He created the concept of the "average man" as the central value about which measurements of a human trait are grouped according to the normal probability curve. In addition to the normal distribution of heights and weights, Quetelet observed that there were relative propensities of specific age groups to commit crimes. He wrote [Mailly, 1875, 179]:

> *What is very remarkable is the frightful regularity with which crimes are repeated. Year after year are recorded the same crimes, in the same order, with the same punishments; in the same proportions. . . . The number condemned to the prison, irons, and the scaffold is as certain as the revenue of the state. We can tell in advance how many individuals will poison their fellows, how many will stain their hands with human blood, how many will be forgers, as surely as we can predict the number of births and of deaths.*

Quetelet's use of the terms and concepts of physics in the study of man and his social systems provoked wide argument and discussion on the issue of "free will versus

determinism." He was strongly convinced that there were discoverable principles dictating man's behavior [Mailly, 1875, 179–180]:

> *Man, without knowing it, and supposing that he acts of his own free will, is governed by certain laws from which he cannot escape. We may say that the human species, considered as a whole, belongs to the order of physical phenomena. . . . Although his will is restrained within very narrow limits, man contains within him moral forces which distinguish him from the animal, and by which he can, to some extent, modify the laws of nature. These perturbing forces act . . . slowly. . . they are analogous to those astronomical variations in the systems of the world which require centuries for their investigation. The study of the natural and perturbing forces of man, in other words, social mechanics, would develop laws as admirable as those which govern celestial and inanimate bodies. . . . If science has advanced thus in the study of worlds, may we not look for equal progress in the study of man? Is it not absurd to suppose that, while all else is controlled by admirable laws, the human race alone is abandoned to blind chance?*

In his 1835 book *On Man and the Development of His Faculties*: *An Essay in Social Physics*, Quetelet criticizes Malthus and the economists who came after him for not clearly establishing the necessary foundation for bringing the theory of population within the domain of the mathematical sciences. He proposes two principles to fill this "important gap" [Quetelet, 1969, 49]. First, population tends to grow according to a geometric progression. Second, "the resistance, or the sum of the obstacles opposed to the unlimited growth of population, increases in proportion to the square of the velocity with which the population tends to increase."

Later Quetelet draws an analogy between the growth of population and the motion of a body through a resisting medium. His writing is somewhat obscure on these points. No mathematical treatment is given and, although he claims to have made "numerous researches" on this subject, none are presented. Quetelet's comments, however, were the probable source that stimulated the first detailed study and presentation of logistic growth. This was the work of Quetelet's Belgian colleague Pierre-François Verhulst.

C. Pierre-François Verhulst

Pierre-François Verhulst. From a portrait by L. Flameng. Copyright Bibliothèque royale Albert Ier, Brussels (Cabinet des Estampes).

Pierre-François Verhulst, born in Brussels on October 28, 1804, was a brilliant student who received his Doctor of Science degree from the University of Ghent after only 3 years of study. His mathematical research included contributions in the calculus of variations, the study of maxima and minima of functions, and number theory, as well as in the applications of probability.

Plagued by poor health, Verhulst traveled to Italy in 1830. While in Rome, he worked for reforms in the government of the pontifical states. Hoping to persuade the Pope to grant a constitution to the residents, Verhulst drew up a proposed pact. It was well received by several foreign ministers, but the confidential document fell out of the hands of the diplomats and into the clutches of the police. Fearing physical attack, Verhulst prepared to barricade himself in his lodgings to withstand a possible siege. He was ordered to leave Rome and return to Belgium.

Back home, he made an unsuccessful attempt to enter politics and tried his hand at writing historical essays. Finally, he returned to the academic life, accepting appointment at the free university of Brussels in 1835. There he taught mathematical subjects ranging from geometry and trigonometry to calculus and probability, as well as astronomy and celestial mechanics. Under the influence of Quetelet, under whom he had studied at Ghent, he investigated the applications of statistical tools to social problems.

Verhulst developed a model of population growth he called "logistic growth"; it is the one studied in Section IV. His memoirs on the subject were published in 1838, 1845, and 1847. Although he attempted to test his model on actual population data, he was frustrated by the inaccurate census information then available. He noted, for example, that the figures on the population of England were obtained from a consideration only of the number of births. These, in turn, were counted by examining the number of babies baptized into the Church of England. Thus, religious dissenters, infant deaths, and immigrants were overlooked. "Probably owing to the fact that Verhulst was greatly in advance of his time, and that the then existing data were quite inadequate to form any effective test of his views, his memoirs fell into oblivion," wrote G. Udney Yule in a presidential address to the British Royal Statistical Society in 1925, "but they are classics on their subject."

Verhulst's discovery of the logistic curve and its application to population growth was forgotten for 80 years. It was rediscovered independently by two American scientists working at Johns Hopkins University, Raymond Pearl and Lowell J. Reed.

D. Raymond Pearl

"It is likely that biology will eventually be as full-panoplied with mathematically expressed theory as physics now is. The process is already started, and the history of the old natural sciences like astronomy, physics and chemistry admits of no doubt as to the final outcome. There is no substitute for mathematics to state in rational shorthand the relations between natural phenomena or generalizations about them."

This was the prediction of the American biologist, geneticist, and statistician Raymond Pearl [1939]. A prolific and articulate writer on many subjects, Pearl was widely respected by his colleagues for his applications of statistics to biology, and he was well known to the public of his day as a provocative commentator on human behavior.

The Pearl family traced its ancestry back to Pearls who entered England at the time of the Norman Conquest in 1066. The first to settle in the United States was John Pearl,

who arrived about 1670. His descendants for the next 200 years remained in the region of New England comprising northeastern Massachusetts, southwestern Maine, and southern New Hampshire. It was in Farmington, New Hampshire, that Raymond Pearl was born on June 3, 1879.

He attended the local elementary and secondary schools and entered Dartmouth College at the age of 16 to pursue—or so his parents and grandparents intended—the study of Greek and Latin. Like many college students today, Pearl was intoxicated at first by the relative freedom and extracurricular activities of the campus. His interests were reflected by low grades in his freshman year. It was in that year, however, that he discovered his true intellectual interest.

Raymond Pearl

Biology was one of the required courses for first-year students at Dartmouth at that time, and it appealed to Pearl. By the end of the first week of classes, he was asking the instructor to help him switch from a classics major to the natural sciences. "The subject obsessed him," according to one commentator. "He talked, thought, studied, and dreamed in terms of biology." Although he was the youngest student in his class at Dartmouth, by his senior year Pearl was serving as assistant in the general biology course.

Upon receiving his bachelor's degree in 1899, Pearl began work in the doctoral program in zoology at the University of Michigan, which he completed in 3 years. After a short period as a zoology instructor at Michigan, he embarked on a 2-year study period at the University of Leipzig, the Marine Biological Station in Naples, and the University of London. In London, he studied biometrics under Karl Pearson, whose influence led Pearl to his life's work on the use of statistics to study populations.

The remainder of Pearl's professional life was spent at the Johns Hopkins University in Baltimore. He served there as professor of biometry and vital statistics in the School of Hygiene and Public Health, professor of biology in the School of Medicine, research professor and director of the Institute of Biological Research, and statistician at Johns Hopkins Hospital. He died of a coronary thrombosis on a weekend trip to Hershey, Pennsylvania, on November 17, 1940.

In a biographical memoir, H. S. Jennings [1943], once Pearl's teacher, wrote of his former student, "He was a man of unusual height and weight, physically an impressive figure. His was a masterful personality, of extraordinary resourcefulness and initiative, of wide knowledge, astonishing power of work, remarkable versatility and scope, and strong ambitions. His interest in biology was encyclopedic. In his contributions he touched upon most aspects of the subject. . . . The breadth of Pearl's interests did not mean that his interest in particular subjects was weak. On the contrary, his interest in any subject to which he gave his attention was so intense that at any given moment he might seem a partisan and propagandist of a particular field or method of biological science."

During his 40-year professional career, Pearl wrote nearly 700 technical articles and essays and 17 books. His work appeared in journals of research in zoology, genetics, physiology, medicine, and statistics, agricultural publications, encyclopedias, newspapers, popular science magazines, literary and political journals. "This is a remarkable record of publication," wrote Jennings. "It may be questioned whether in America it has ever been equaled by a man of science, in extent and variety."

The range of Pearl's work can be seen by comparing his first paper, "On preparing earthworms for section," which appeared in the *Journal of Applied Microscopy* in 1900 and his last, "Some biological considerations about war," written for the *American Journal of Sociology* in 1940.

A glance through the 37-page list of his publications shows works on animal behavior, genetics, care and breeding of poultry, laboratory and field techniques in biology, theoretical and practical results in statistics, disease, longevity and mortality, contraception, eugenics, world overpopulation, business cycles, food prices, religion and Darwinism, and philosophical pragmatism.

It was the biology of man, however, that attracted more of Pearl's activity than any other subject. Many of his research results in this area prompted controversy and were widely discussed. An extensive statistical study in 1926, for example, of the effects of the use of alcohol on longevity and mortality persuaded Pearl that moderate consumption of alcohol is not harmful. A similar study in 1938 convinced him that tobacco is harmful to human life even in small quantities. Other research led Pearl to conclude that length of life varied inversely with the pace of living, that intellectuals had a better chance than manual workers of living longer, and that heredity dominated over environment in influencing many important parameters of life.

Socially prominent and popular, Pearl was famous for his excellent dinner parties. He was a connoisseur of good food and wine and possessed, according to one witness, "an almost boyish delight in playing at times the role par excellence himself of amateur cook and salad mixer."

One of Pearl's strongest recreational interests was music. He led for some years an evening amateur music ensemble. When a report on the Dartmouth class of 1899 was written 35 years after its graduation, Pearl's devotion to music was recalled [Dartmouth College, 1941]:

> *He might be the first American to deliver the Heath Clark lectures at the University of London, or the most skillful juggler of the logistic curves of Verhulst; to us he was still the boy cornetist and the fellow who single-handed conjured the first Dartmouth band into existence out of rustic young neophytes and rusty and discarded tubas. He was our full-fledged impresario before we even knew there was such a word, and no crowd of urchins ever followed the Pied Piper of Hamelin so devotedly and gaily as we of '99 and all our Dartmouth contemporaries followed*

the imperious form of Pearl, as in corduroy or white duck trousers and with much "windy suspiration of forced breath" he poured strange harmonies on the campus air.

Pearl served as president of several important scientific organizations, including the American Society of Zoologists, American Society of Naturalists, American Statistical Association, American Association of Physical Anthropologists, and International Union for Scientific Investigation of Population Problems. He received many honorary degrees and other awards, including membership in the National Academy of Sciences.

Pearl's own views toward the future of man may be gleaned from the final sentences of his last published work. "The standard pattern of national behavior, to which there are no exceptions, is to combat evil with evil," he wrote [Pearl, 1941]. "But real and enduring peace will never be achieved by such techniques. For a true evolution of new patterns of sociality that will be lasting and embrace all mankind there must first evolve among men more decency and dignity, more tolerance and forbearance, and more capacity of co-operation for the common good in the conduct of human life. The prime condition necessary for the meek to inherit the earth is that they shall *abound* in the qualities of meekness."

E. Lowell Jacob Reed

Raymond Pearl was perhaps best known to the public for the projections of U.S. population size that were the product of research completed with his colleague Lowell J. Reed. Like Pearl, Reed's family roots were in New England. Born in Berlin, New Hampshire, on January 8, 1886, he received his bachelor's and master's degrees from the University of Maine. After completing his Ph.D. in mathematics at the University of Pennsylvania, he returned to Maine to teach physics and mathematics.

His academic career was interrupted for a period of service during World War I in Washington as chief of the Bureau of Tabulations and Statistics for the War Trade Board. His ties with the government continued in later years as he served as a consultant for the Army, Navy, Air Force, Selective Service, and Veterans Administration.

Lowell Reed at his New Hampshire farm.

Reed's long association with Johns Hopkins began in 1918 when he was appointed an associate professor of biostatistics in the School of Hygiene and Public Health. Several years later, he succeeded Pearl as chairman of the department. An effective administrator, Reed served as dean of the public health school, vice president of Johns Hopkins Hospital, and vice president of the university.

His principal research interests were in the fields of mathematical methods in biology and medicine, mathematical statistics, and demography. He was internationally known for contributions to biostatistics and public health administration. His work included many important advances in the study of epidemics.

In the early 1920s, Pearl and Reed worked on interpolation formulas for population curves, with special reference to the United States. After trying various purely empirical curve-fitting equations, they realized that no such formula could be regarded as a general law of population growth, however good it might prove for practical purposes over a limited period. Consideration of the general principles underlying population changes led them to the mathematical model of logistic growth. Pearl and Reed's discovery of the logistic curve was quite independent of Verhulst; they learned of the Belgian's work months after deriving all the mathematical details for themselves.

At a 1925 conference, Reed predicted that it would take a century for the United States to reach a population of 200 million, and that when it did, there would be such pressure on the country's food resources that new sources of sustenance would have to be found in the tropics.

Although this prediction turned out to be wrong, others that he made, such as forecasting the rapid growth of the metropolitan New York region, were both accurate and useful to planners.

At the age of 67, Reed retired from Johns Hopkins after 35 years of service. He hoped to return to his 300-acre farm in Shelburne, New Hampshire to enjoy the quiet life of its rugged woodlands. Barely 3 weeks after his retirement began, Reed was called back to Johns Hopkins and asked to accept the presidency of the university. He served well in this position for 3 years and finally found permanent retirement at 70.

Happily turning over his office to Milton Eisenhower, younger brother of President Dwight Eisenhower, Reed moved back to his native New Hampshire. He died in the town of his birth on April 29, 1966.

F. Evelyn C. Pielou

Considered by many to be the creator of the field of Mathematical Ecology, Chris Pielou (1924–) has firm views on the centrality of mathematics in the sciences. Addressing a graduating class of mathematics students at the University of British Columbia in 1991, she stated:

> *I'm sure people will have pointed out to you (or, possibly,* you *have pointed it out to* them*) that "Mathematics is the Queen of the Sciences," just as Gauss proclaimed. Of all possible subjects of study, it is the one which commands the most awe and veneration. But before you get too carried away with this thought, remember that math is also the servant of science, except for parts of some sciences such as paleontology that haven't become mathematical yet. They will. All scientists depend on a mixture of experiments and observations to do their work. The next*

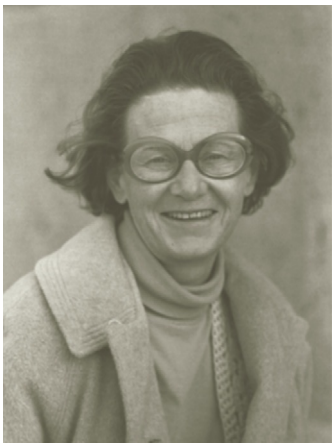
stage, theoretical development, entails logical, mathematical argument. More than a hundred years ago, Lord Kelvin (he of the absolute temperature scale) put it thus:

"When you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind; it may be the beginning of knowledge but you have scarcely, in your thoughts, advanced to the stage of science."

In brief, most scientific data don't become scientific until they have been put into numerical form. I believe Kelvin should have gone farther than that: to my mind, he should have added that scientific notions cannot become part of science until they've been expressed as mathematical equations (or occasionally, as in*equalities). Until this happens, notions aren't hypotheses—they're just hunches.*

I believe the majority of non-scientists are unaware of this, of the dependence of science on math. This may explain why so many people say, complacently, "Of course, I'm lousy at math but . . . " and then go on to imply that their mental powers are perfect apart from this trivial defect. Well, it isn't trivial—a person who blocks out math is a mental couch potato. You, by mastering the queen of the sciences, become truly "fit" in the world of the mind. Whether you are a pure mathematician, or an applied mathematician, or a theoretical statistician, or a computer scientist, you are each of you a mental athlete. The possessor, and user, of a rigorously logical mind deserves public recognition and admiration just as much as a champion athlete or a well-known movie star—but you wouldn't think so to read the newspapers.

Eveyln Chrystalla Pielou was born in Bognor Regis, England, on February 24, 1924. She earned an honors degree in botany from the University of London in 1950, where she later obtained her doctoral degree in Statistical Ecology in 1962. In addition, she earned a Senior Doctorate from the university in 1975.



Reproduced with permission of Professor Pielou

Evelyn C. Pielou

First employed as a research scientist in Canada's Federal Departments of Forestry and Agriculture, Dr. Pielou then spent 1-year term as a visiting professor at North Carolina State University and Yale University. She moved to Queen's University in Kingston,

Ontario, where she was appointed a professor in the Biology Department from 1968 to 1971. A move eastward took her to Dalhousie University in Halifax where she served as a Killam Research Professor and professor of Biology for the next 10 years. Dr. Pielou spent the last 5 years, prior to retiring in B.C. in 1986, working as a research professor in the Biology Department of the University of Lethbridge.

In addition to authoring more than 60 research papers, Pielou wrote a number of books on a variety of subjects. These include *An Introduction to Mathematical Ecology*, *Population and Community Ecology*, *Ecological Diversity*, *Mathematical Ecology*, *Biogeography*, *The World of Northern Evergreens*, *After the Ice Age*: *The Return of Life to Glaciated North America*, *A Naturalist's Guide to the Arctic*, *Fresh Water*, and *The Energy of Nature*.

Many universities and scientific societies have recognized Pielou's contributions to the mathematical modeling of natural systems. The Canadian Botanical Association awarded her the George Lawson Medal in 1984. The Ecological Society of America presented her the Eminent Ecologist Award in 1986. In 1990, she received the Distinguished Statistical Ecologist Award from the International Congress of Ecology. Several universities have given her honorary degrees, and she has been elected to the Royal Society of Arts.

G. A Final Note of Caution

In discussing the difficulties and limitations of the logistic model for the growth of human populations, the sociologist Donald Olen Cogwill wrote [Bose et al. 1970]:

> Initially the theory was based upon experiments with yeast, fruit flies, and chickens and the conditions of these experiments should be carefully noted:
>
> **1.** The initial population was very small in relation to the space that was provided for it;
>
> **2.** Ample food was provided throughout the experiments;
>
> **3.** The food was introduced into the experimental environment by the experimenter and it was not generated by the species that was the subject of the experiment; and
>
> **4.** The spatial limits of the environment were held constant.

The reader should consider carefully whether such conditions are reasonable to assume operative for human populations.

EXERCISES

II. The Pure Birth Process

1. The rate of growth of a certain population of bacteria in a culture is directly proportional to the size of the population. If an experiment begins with 1,000 bacteria and one hour later the count is 1,500 bacteria, then how many bacteria are present at the end of 24 hours?

2. Suppose that 20 years ago the population of a town was 2,000, and that the population increased continuously at a rate proportional to the existing population. If the population of the town is now 6,000, find a formula relating population and time. What has been the rate of growth?

3. Suppose the population of a city doubles its original size in 50 years and triples it in 100 years. Can the population be increasing at a rate proportional to the number present? Why, or why not?

4. Suppose the population of a yeast colony is given by $P(t) = P_0 e^{at}$ and the population at time t_1 is P_1. Find a formula for a in terms of t_1, P_0, and P_1.

5. If population $P(t)$ is growing exponentially, prove that the changes in P in successive time intervals of equal duration form the terms of a geometric progression. This is the source of Thomas Malthus's famous dictum: "Population, when unchecked, increases in a geometrical ratio. Subsistence increases only in an arithmetic ratio. A slight acquaintance with numbers will shew the immensity of the first power in comparison of the second."

6. If a certain population increases at a rate proportional to the number in the population and doubles in 45 years, in how many years is it multiplied by a factor of 3?

7. A population of bacteria grows exponentially. When initially observed, there were 100,000 bacteria. Another observation t_1 minutes later showed 200,000 bacteria. A third observation was taken 10 minutes after the second one; this time 1,000,000 bacteria were present.

(a) Find the equation of growth of the bacteria.

(b) How many bacteria were there after 20 minutes?

(c) What is the value of t_1?

8. If the population of a country is undergoing exponential growth at a rate of r percent per year, show that the population doubles every $(\log 2)/r$ years. This number is called the "doubling time." Compute the doubling time if $r = 2$.

9. (*Emmell*) A human birth rate of 50 live births annually per 1,000 population is considered very high. In 1971, several countries in Africa had birth rates of 52 per 1,000. A low birth rate today is about 15 per 1,000; in 1971, Sweden and Luxembourg had the world's lowest birth rates of 13.5 per 1,000. Current death rates range from 5 deaths per 1,000 to 30 per 1,000. On a worldwide basis, the annual birth rate in 1971 was 34 per 1,000 and annual death rate was 14 per 1,000. What is the annual rate of increase of the world's population? Complete the following table using the results of Exercise 8.

Country	Growth rate per 1,000	Doubling time
East Germany	.1	
Denmark	.5	
United States, Japan	1.1	
Argentina, World	1.5	
Afghanistan	2.5	
Ghana	3	
Costa Rica	4	
Kuwait	8.2	

10. What happened in the United States between 1860 and 1870 that could have accounted for a halt in exponential growth?

11. Assume that the U.S. population has grown exponentially. Estimate the growth rate using each of the following years in place of the year 1830 as done in the text. Compare each set of predictions with the actual data.

(a) 1800

(b) 1850

(c) 1900

(d) 1970

12. In the pure birth process, suppose the birth rate is not constant, but instead is proportional to P^k for some small positive constant k. Find the differential equation for growth of a population fitting this description. Solve the equation and interpret the result. This model gives a good picture of the population growth in some developing countries (Watt, K. E. F., *Ecology and Resource Management*: A *Quantitative Approach*, New York: McGraw-Hill, 1968).

13. *San Francisco Chronicle* columnist Herb Caen observed (October 27, 1993): "When Elvis Presley died in 1977, there were an estimated 37 Elvis impersonators in the world. By 1993, there were 48,000 Elvis impersonators, an exponential increase. Extrapolating from this, by 2010 there will be 2.5 billion Elvis impersonators. The population of the world will be 7.5 billion by 2010. Every 3rd person will be an Elvis impersonator by 2010."

(a) If the data for 1977 and 1993 are correct and the population of Elvis impersonators did in fact grow exponentially, what was the annual growth rate a?

(b) Using this value for a, determine what the impersonator population would actually be in 2010 if the population continues to grow at the same exponential pace. Is 2.5 billion correct?

(c) In what year would the number of Elvis impersonators reach 2.5 billion?

(d) Is there some year when there would be more Elvis impersonators than people?

14. The population dynamics of a city may also be substantially influenced if there is a large rate of individuals moving into the community (immigration) or moving away (emigration). A simple model of exponential growth with immigration is $\frac{dP}{dt} = aP + b$ where a and b are positive constants.

(a) Show that the change of variable $y(t) = P(t) + b/a$ transforms the model into a pure exponential growth model: $dy/dt = ay$.

(b) Find $P(t)$ as an explicit function of t.

15. During periods of war or great civil unrest, nations may experience significant declines in population due both to higher death rates and to large scale emigration. Analyze the model $\frac{dP}{dt} = aP + b$, where b is a negative constant both in the case in which a is positive and in which a is negative. What can you conclude about the long-term prospects for the population in these situations? Locate, if possible, population data for a country experiencing large emigration (e.g., Rwanda in the 1990s, Iraq and Syria in the first decades of the 21st century) and determine whether this model provides an accurate picture of the observed data.

III. Exponential Decay

16. A certain radioactive substance has a half-life of 10 years. What fraction of an amount of this substance decays in 15 years?

17. Verify the claim made in the text that the half-life of a radioactive substance is independent of P_0.

18. A carved wooden stick found at an archaeological site near Madison, Wisconsin, had 40 percent of the radioactivity of a living tree. When was the stick carved?

19. The Shroud of Turin, which many people believe was used to wrap Christ's body, bears detailed front and back images of a man who appears to have suffered whipping and crucifixion. First displayed in France in the 1350s, the shroud was brought in 1578 to Turin, where it was placed in the royal chapel of Turin Cathedral in a specially designed shrine. Tests done on the Shroud of Turin in 1989 found that it contained 92% of its original fraction of Carbon-14. If the half-life of Carbon-14 is 5,530 years, estimate the true age of the shroud.

20. A population, initially of 10,000 individuals, has an annual decay rate of .1. In how many years will the population decrease to 1 person?

21. *Newton's Law of Cooling* asserts that the rate at which an object cools is proportional to the difference between the temperature of the object and the temperature of the environment in which the object is immersed. One proposed application of Newton's Law of Cooling is the determination of the time of death of a murder victim whose body is found in a hotel room.

(a) If V denotes the temperature of the object (victim) at time t, show that Newton's Law of Cooling may be formulated as a differential equation

$$dV/dt = k(V - R)$$

where R is the constant temperature of the room and k is a negative proportionality constant dependent on the thermal properties of the corpse.

(b) Verify by substitution that $V(t) = R + Ce^{kt}$ is a solution to the differential equation.

(c) Derive the solution to the differential equation either (1) by separating the variables and integrating directly or (2) by making the substitution $y(t) = V(t) - R$.

(d) Determine the value of the constant C if the victim's temperature is V_0 at time $t_0 = 0$.

(e) If the temperature of the victim is V_1 at some positive time t_1, then show that k has the value

$$\frac{1}{t_1} \mathit{In} \frac{V_1 - R}{V_0 - R}$$

(f) From the solution $V(t) = R + Ce^{kt}$, show that if the victim has a temperature of V_N at time t_N, then t_N can be found as

$$t_N = \frac{1}{k}\, ln\, \frac{V_N - R}{C}$$

(g) Police discover the dead body of a woman inside her Los Angeles condominium around midnight. The medical examiner arrived on the scene at 12:30 A.M. and immediately took the murder victim's body temperature; it registered as 94.6°F. The police investigation of the murder scene lasted one hour. Just before the body was wheeled away and the police left, the examiner took the victim's temperature again; this time, his thermometer showed 93.4°F. If the room was maintained at a temperature of 68°F, estimate the time of death.

22. A forest products company cuts down 5% of the trees in a national forest area each year, but also plants 1,000 new trees on an annual basis. What happens to the number of trees in the area over the long term? Find explicit expressions for the tree population as functions of time using a continuous model and a discrete model. Graph these functions if the initial tree population was (a) 50,000 and (b) 10,000.

23. In what ways is the forest problem of Exercise 22 similar to the credit card problem of Chapter 1?

IV. Logistic Population Growth

24. A third derivation of the Verhulst-Pearl equation is based on the notion that a term involving the square of the population is a reasonable measure of "crowdedness." It would represent the frequency with which members of the population encounter each other. Is it reasonable that this frequency would have an inhibitory effect on the rate of population growth? Why? How does this lead to the logistic equation?

25. Models of population growth may be derived from the differential equation $dP/dt = f(P)$ by various choices of simple functions for f. For each of the following types of functions, determine reasonable choices for the signs of the coefficients, solve the resulting equations and interpret the results:

(a) $f(P) = a$

(b) $f(P) = a + bP$

(c) $f(P) = a + bP + cP^2$

(d) $f(P) = a\sin(bt + c)$

26. At some point in the solution of the logistic differential equation, we implicitly assumed that population P was always below the carrying capacity a/b. Where? How valid is this assumption?

27. Suppose a forest fire destroys a large portion of the resources on which a population feeds. The carrying capacity of the environment is then below the initial population P_0. Analyze the logistic model in this situation to the point that you can sketch the graph of population as a function of time. (Compare with Exercise 26.)

28. Show that the logistic curve has a single point of inflection. At what value of t does it occur? What is the corresponding population? How does it compare to the limiting population? Is the logistic curve symmetric about the point of inflection?

29. Show that some of the answers for Exercise 28 can be obtained from the Verhulst-Pearl equation without solving for P in terms of t.

30. In Pearl and Reed's model of U.S. population growth, find the year when the rate of population growth first began to slow (see Exercise 28).

31. How can you determine the constants in the logistic equation if the populations at three different times are known but the times are not equally spaced?

32. Assume that the growth of population in the United States from 1790 to 1860 is adequately explained by a pure birth process. How closely does the logistic model explain growth in population from 1860 to 1970? Take 1870, 1920, and 1970 as the years to use in computing the constants in the logistic equation. What does this model predict as the "carrying capacity" of the United States?

33. Find the value of the constants in the equation of the logistic curve using the census data for the years 1790, 1880, and 1970. What is the predicted carrying capacity? How well does the resulting curve fit actual census data?

34. An initial population of 100 inhabits an area with a carrying capacity of 100,000. In the first year, the population increases to 120. Assume that the population follows logistic growth.

(a) Determine the population as an explicit function of time.

(b) How many years will it take the population to reach 95,000?

35. Find census data for the population of a western European nation and determine how valid the logistic model is for that population. Take t_0, t_1, and t_2 to be spaced 100 years apart.

36. Repeat Exercise 34 for world population. How accurate do you think the available census data are?

37. Sociologists recognize a phenomenon called "social diffusion," the spreading of a piece of information, a technological innovation, or a cultural fad among a population. The individuals in the population can be divided into those who have the information and those who do not. In a fixed population whose size is known, it is reasonable to assume that the rate of diffusion is proportional to the number who have the information and the number yet to receive it.

(a) If x denotes the number of individuals in a population of N people who have the information, then show that a mathematical model for social diffusion is $dx/dt = kx(N - x)$, where t represents time and k is a proportionality constant.

(b) Solve the equation in (a), and show that it leads to a logistic curve.

(c) At what time is the information spreading fastest?

(d) How many people will eventually receive the information?

(e) Discuss how this model might be modified to analyze an epidemic of a communicable disease.

38. Show that if $a = 2.57$, then as k increases, the values of Q_k in the discrete logistic model cycle through 16 values.

V. The Allee Effect

39. The Allee effect has been described as the ecological equivalent of the maxim "The more the merrier." In what sense is this statement true?

40. The existence of the Allee effect has been used to justify the claim that "the evolution of social structure was not only driven by competition, but that cooperation was another, if not the most, fundamental principle in animal species." How are competition and cooperation reflected in our model?

41. Verify that the signs of dP/dt in Fig. 3.18 are correct.

42. For the logistic model with Allee effect, show both of the following:

(a) K is a stable equilibrium.

(b) $c > 0$ is an unstable equilibrium.

43. Show that the per capita growth rate in the logistic model with Allee effect reaches its maximum value at $P = (K + c)/2$.

44. Show the growth rate in the logistic model with Allee effect, there is an inflection point at $P = (K + c)/3$.

45. One alternative model with the Allee effect is the differential equation

$$\frac{dP}{dt} = bP\left(1 - \frac{P}{K}\right)\left(1 - \frac{c + A}{P + A}\right)$$

where c is the Allee threshold and K is the carrying capacity. Show the following:

(a) The Allee effect is absent if and only if $c = -A$.

(b) B is the maximum per capita growth rate in the absence of the Allee effect.

(c) The constant A affects the overall shape of the per capita growth rate curve in the sense that the curve becomes flatter as A increases and reaches a lower maximum value.

(d) The Allee effect is strong if c is positive.

SUGGESTED PROJECTS

1. Consider a model for the population of scientists alive at any given time. It has been reported that 90% of all scientists who have ever lived are alive today. What sort of model is consistent with this fact?

2. A simple generalization of the logistic equation is the differential equation $dP/dt = aP + bP^2 + cP^3$. Analyze the consequences of this model. Discuss how to evaluate the constants a, b, and c. Does this model give a good picture of population growth in the United States from 1790 to the present?

3. Some animal populations are periodically reduced by hunters or trappers for commercial gain. Consider the problem of determining the optimal rate of removal by a hunter who wishes to maximize *long-range* economic gain; killing the entire population in 1 year means great profits that year but no income in

subsequent years. Determine effective strategies for the hunter if the population is growing (a) exponentially and (b) logistically. See the papers of Colin W. Clark (References) for some suggested approaches.

4. The *average* (or *per capita*) *growth rate* of a population is given by $\frac{1}{P}\frac{dP}{dt}$. In the logistic model, this average growth rate is largest when the population is smallest. (Why?) This is an unrealistic model for some species that may face extinction if the population becomes too small. Suppose that c is the minimum viable population for such a species. Consider the modified logistic equation $dP/dt = (a - bP)(P - c)$. Solve this equation and interpret the results. In particular, show that the population eventually becomes extinct if P is ever less than c.

5. (*Grossman and Turner*) Biologists have discovered that the growth, survival, and reproduction of cells are determined by nutrients flowing across the cell walls. During the early stages of a cell's growth, the rate of increase of the weight W of the cell will then be proportional to its surface area. If the shape and density of the cell do not change during growth, the weight will be proportional to the cube of a radius while the surface area is proportional to the square of a radius. Show that a reasonable model for the growth of the weight of the cell as a function of time is given by the solution of the differential equation $dW/dt = cW^{2/3}$ where c is a positive constant. Investigate the consequences of this model. What are the limitations of this model of cell growth? Develop a differential equation model that takes into account the fact that there may be a maximum weight that the cell cannot exceed.

6. In experiments at Columbia University's Institute of Cancer Research, Fred R. Kramer and his associates studied the growth of an RNA population in the presence of a fixed concentration of replicase molecules. Kramer observed that the early stage of growth is nearly exponential $(dP/dt = aP)$, but that after a certain period of time, population approaches linear growth $(dP/dt = b)$. Develop a differential equation model for population growth consistent with these observations. Solve the equation and interpret the results. Derive some predictions from the model that Kramer can test against his other experimental data.

7. Investigate discrete models that include logistic growth with an Allee effect and examine conditions under which chaos may arise. The paper by Elaydi and Sacker cited in the online references is a good starting point.

BIOGRAPHICAL REFERENCES

T. R. Malthus

Malthus, Thomas R., *An Essay on the Principle of Population and a Summary View of the Principle of Population*, intro. by Antony Flew, ed., Baltimore: Penguin, 1970.

R. Pearl

"A Thirty-Fifth Report of the Class of 1899 of Dartmouth College," Hanover, NH: Dartmouth College, 1941.

Jennings, H. S., "Biographical Memoir of Raymond Pearl," *National Academy of Science. Biographical Memoirs* **22** (1943): 295–347.

Pearl, Raymond, "Review of *Mathematical Biophysics* by N. Rashevsky," *Bulletin of the American Mathematical Society* (1939): 223–224.

Pearl, Raymond, "Some Biological Considerations about War," *American Journal of Sociology* **46** (1941): 487–503.

L. A. J. Quetelet

Mailly, Edouard, "*Eulogy on Quetelet*," *Annual Report of the Board of Regents of the Smithsonian Institution for the Year 1874*, Washington: Government Printing Office, 1875, 169–183.

Quetelet, Lambert A. J., *A Treatise on Man and the Development of His Faculties*, Gainesville, FL: Scholars' Facsimiles and Reprints, 1969.

P. F. Verhulst

De Seyn, E., ed., *Dictionnaire Biographique des Sciences, des Lettres et des Arts en Belgique*,Tome Second, Brussels: Editions L'Avenir, 1936, 1128.

Miner, John R., "Pierre-François Verhulst, the discoverer of the logistic curve," *Human Biology* **5** (1933): 673–685.

W. C. Allee

Karl Schmidt, "Warder Allee, 1885–1955," *Biographical Memoirs, National Academy of Sciences* 30 (1957), www.nasonline.org/publications/biographical-memoirs/memoir-pdfs/allee-warder.pdf.

You can find a listing of references and suggestions for additional reading on the book's website, www.wiley.com/college/olinick

CHAPTER 4 Ecological Models: Interacting Species

There are craft standards in both mathematics and ecology and the ideal interdisciplinary study simultaneously enhances our understanding of the empirical world and constitutes an example of elegant craftsmanship by both ecological and mathematical standards. That is a difficult set of criteria, but there is no reason to believe that science at its best is easy.

—Lawrence B. Slobodkin

I. Introduction

The previous chapter developed some models for population growth of a single species inhabiting an environment in which the amount of resources never changed and the numbers of other species also remained fixed. Although such a situation may sometimes be approached in laboratory experiments, an effective mathematical model should not ignore the fluctuations of the other important variables in an ecosystem.

In this chapter we present several models that attempt to represent the population dynamics that can occur in a system when two or more species interact with each other in the same environment. As is the case with most of the material in this book, we only consider relatively simple models. We will look in detail at two particular models that were the classic beginnings of mathematical ecology. They form the bases on which scientists construct more sophisticated models.

Those readers unfamiliar with partial derivatives and the other basic ideas of the calculus of several variables should read Appendix IV before tackling Section III.

II. Two Real-World Situations

A. Predator and Prey

Consider first the effects of interdependence of two species, one of which serves as food for the other. A classic formulation of this situation is that of a population of gazelles, who feed

only on grass, and a population of leopards, who feed exclusively on gazelles. The assumptions usually made about the situation are these:

1. In isolation, the rate of change of population of one species is proportional to the population of that species. In the absence of leopards, we assume that the gazelle population will exhibit exponential growth. If there are no gazelles, the leopard population will undergo a pure death process.
2. There is always so much grass that the gazelles have an ample supply of food. The only food available to the predatory leopards are the gazelles.
3. The number of kills of gazelles by leopards is proportional to the frequency of encounters between the two species. This, in turn, is proportional to the product of the populations of gazelles and leopards. Thus, there will be few kills if there are few gazelles or few leopards, and many kills only when both populations are relatively large.

If G denotes the population of gazelles at time t and L is the number of leopards, then the predator-prey model asserts that G and L are functions of time that satisfy the pair of first-order differential equations

$$\frac{dG}{dt} = aG - bGL$$

$$\frac{dL}{dt} = mGL - nL$$

where a, b, m, and n are positive constants.

Although this model was initially developed to study actual animal populations, it has been used to consider other interactions as well. In a series of research papers, George Bell applied the concepts of this model to analyze the immune response to infections. When a living being is infected by a replicating organism, such as bacteria or a virus, an immune response may be produced. The response is characterized by the production of antibodies that bind to the infecting material and hasten its destruction. Antigen plays the role of prey (gazelles) and antibody the role of predator (leopard) [Bell, 1973]. Other scholars have used variants of this model to study interactions between workers and capitalists [Goodwin, 1967] and between humans and vampires [Hartl et al. 1992]. Predator-prey analogies have also been pursued to gain better understanding of the spread of epidemics and of revolutions [Epstein, 1997].

B. Competitive Hunters

A different situation involving interacting populations is one in which two species have a common prey or food source. Here the predators are in competition with each other. Each removes from the environment a resource that would stimulate the growth of the population of the other. We shall refer to this situation as one involving *competitive hunters*.

The assumptions about this situation are somewhat similar to the ones set down in the predator-prey case:

1. In the absence of one of the predators, the other predator's population increases at a rate proportional to its size.
2. There are sufficient numbers of prey to sustain any level of predator population.
3. The competition between the predators is proportional to the product of the population of these two species.

If U and V denote the populations of the two predators, then the model asserts that U and V are functions of time t satisfying the pair of differential equations

$$\frac{dU}{dt} = aU - bUV$$

$$\frac{dV}{dt} = mV - nUV$$

where a, b, m, and n again are positive constants.

III. Autonomous Systems

A. Three Autonomous Systems

In the discussion of mathematical models in this text, we have presented three different pairs of differential equations: one for Richardson's arms race between two nations, one for a predator-prey relationship, and one for a competitive hunters situation. There are certain similarities in these systems we want to explore.

In each of the three systems of differential equations, there are two variables, call them x and y, which are functions of a third variable, say t. In each case, the model is an assertion that a certain pair of differential equations involving these variables is true. The models look like this:

$$\text{Arms Race:} \quad \frac{dx}{dt} = ay - mx + r \quad \frac{dy}{dt} = bx - ny + s$$

$$\text{Predator-prey:} \quad \frac{dx}{dt} = ax - bxy \quad \frac{dy}{dt} = mxy - ny$$

$$\text{Hunters:} \quad \frac{dx}{dt} = ax - bxy \quad \frac{dy}{dt} = my - nxy$$

In all three models, the differential equations are of the type

$$\frac{dx}{dt} = x'(t) = F(x,y)$$

$$\frac{dy}{dt} = y'(t) = G(x,y)$$

The rates of change of x and y are given as explicit functions of x and y alone and do not include the third variable t.

Such systems of differential equations are called *autonomous systems*. A solution of such a system is a pair of scalar functions $x = x(t)$ and $y = y(t)$ such that

$$x'(t) = F(x(t), y(t)) \quad \text{and} \quad y'(t) = G(x(t), y(t))$$

for all t in some interval.

A nonautonomous system would have the form

$$\frac{dy}{dt} = H(x, y, t)$$

$$\frac{dy}{dt} = I(x, y, t)$$

where H and I are functions of the three variables. One such example would be the system

$$\frac{dx}{dt} = xy - 2x + \sin t$$

$$\frac{dy}{dt} = \frac{x}{t} + y^3$$

B. Some Mathematical Facts

If the functions F and G of an autonomous system and their first-order partial derivatives are continuous in some domain D of the xy-plane, then the system always has a solution. Furthermore, if (x_0,y_0) is any point in D and t_0 is any number, then there is a *unique* solution defined on some interval about t_0 satisfying the initial conditions $x(t_0) = x_0$, $y(t_0) = y_0$. (Any advanced level differential equations text will contain a precise formulation and proof of this existence-uniqueness result; in particular, see the books by Hirsch, Smale and Devaney, or Hubbard and West listed in the References.)

It is a simple matter to check whether in our three models the functions F, G, F_x, F_y, G_x, and G_y are continuous over the entire xy-plane. We leave this as an exercise.

Autonomous systems of differential equations have been extensively studied and there is a rich literature about the nature of solutions to such systems. We will consider only a few basic properties here.

As time t varies, a solution $x = x(t)$, $y = y(t)$ of the system describes parametrically a curve lying in the xy-plane. This curve is called an *orbit*, or *trajectory*, of the system. Fig. 2.1 shows two possible orbits for the elementary spiraling arms race model $dx/dt = ay$, $dy/dt = bx$.

In ecological models the concern is primarily with the possible values attained by x and y and only secondarily with the times at which these values are achieved. What is wanted, then, is information about the geometric nature of the possible orbits. The first theorem we have asserts that the orbit is independent of the starting time.

THEOREM 1 If $x = x(t), y = y(t)$ is a solution of the autonomous system

$$\begin{aligned} x'(t) &= F(x, y) \\ y'(t) &= G(x, y) \end{aligned}$$

and t_0 is any constant, then the functions $x_1(t) = x(t + t_0), y_1(t) = y(t + t_0)$ also give a solution to the system.

Proof of Theorem 1 We must show that

$$x_1'(t) = F(x_1, y_1) \quad \text{and} \quad y_1'(t) = G(x_1, y_1)$$

We can easily do this by making use of the chain rule for differentiation. According to the chain rule,

$$x_1' = x'(t + t_0)(t + t_0)' = x'(t + t_0)(1) = x'(t + t_0)$$

and, similarly,

$$y_1'(t) = y'(t + t_0)$$

Since $x'(t) = F(x(t), y(t))$ and $y'(t) = G(x(t), y(t))$, replacing t by $t + t_0$, gives

$$x_1'(t) = x'(t + t_0) = F(x(t + t_0), y(t + t_0)) = F(x_1(t), y_1(t))$$

and

$$y_1'(t) = x'(t + t_0) = G(x(t + t_0), y(t + t_0)) = G(x_1(t), y_1(t))$$

which was to be shown. This concludes the proof. ◇

It is very important to notice the basic distinction between a *solution* of the system and an *orbit* of the system. An orbit is a curve that may be represented parametrically by more than one solution. The pairs of functions $x(t), y(t)$ and $x(t + t_0), y(t + t_0)$, for $t_0 \neq 0$, represent distinct solutions, but they represent the same curve parametrically—that is, both solutions give rise to the same orbit.

Example

The pairs $x(t) = \cos t$, $y(t) = \sin t$, and $x(t) = \cos(t + \pi/3)$, $y(t) = \sin(t + \pi/3)$ are different solutions to the system $x'(t) = -y(t), y'(t) = x(t)$. Both, however, represent the same orbit, the familiar unit circle with equation $x^2 + y^2 = 1$.

Our second theorem guarantees that two distinct orbits for an autonomous system cannot cross anywhere; otherwise there would be two different orbits through the same point.

THEOREM 2 Through any point there passes at most one orbit.

Proof of Theorem 2 Suppose, to the contrary, that C_1 and C_2 are distinct orbits that both pass through the same point (x_0, y_0). Let $x_1(t)$, $y_1(t)$ be a solution that represents C_1 parametrically, and let $x_2(t)$, $y_2(t)$ be a solution representing the orbit C_2.

The two orbits must reach the common point (x_0, y_0) at different times, since otherwise the uniqueness of the solutions would be violated. Thus there are distinct numbers t_1 and t_2 such that

$$(x_1(t_1), y_1(t_1)) = (x_2(t_2), y_2(t_2)) = (x_0, y_0)$$

By Theorem 1, the pair of functions

$$x(t) = x_1(t + t_1 - t_2), \quad y(t) = y_1(t + t_1 - t_2)$$

also serve as a solution to the autonomous system of differential equations.

Note now that $x(t_2) = x_1(t_2 + t_1 - t_2) = x_1(t_1) = x_0$, and, similarly, $y(t_2) = y_0$. By the uniqueness of solutions of the system with prescribed initial values, the pair $x(t)$, $y(t)$ is identical to the pair $x_2(t)$, $y_2(t)$. Thus, the orbit associated with $x(t)$, $y(t)$ must be C_2. On the other hand, from the definition of $x(t)$, $y(t)$, we see that this pair is a parameterization of the orbit given by $x_1(t)$, $y_1(t)$. Hence, the orbit associated with $x(t)$, $y(t)$ must be C_1. The conclusion is that C_1 and C_2 coincide and are not distinct. This contradiction to the initial assumption that the orbits were distinct establishes the truth of the theorem. ◇

Armed with these two theorems and the existence-uniqueness result, we will be able to show that the orbits of an autonomous system must either be single points or "simple" curves.

C. Types of Orbits

Consider the autonomous system of differential equations

$$\frac{dx}{dt} = 7y - 4x - 13$$

$$\frac{dy}{dt} = 2x - 5y + 11$$

One simple solution to this system is the pair of constant functions

$$x(t) = 2$$

for all t

$$y(t) = 3$$

The orbit of this solution is the single point (2, 3) in the xy-plane. A quick calculation shows that $dx/dt = dy/dt = 0$ at this point. This result is consistent with the fact that the derivatives of constant functions are zero.

More generally, suppose there is a constant solution $x(t) = x_0, y(t) = y_0$, for $-\infty < t < \infty$, to an autonomous system. By the uniqueness of solution, no other orbit could pass through the point(x_0, y_0). Since these are constant functions, it is true that

$$x'(t) = 0, \quad y'(t) = 0, \quad -\infty < t < \infty$$

and since the functions are solutions to the system, we have

$$\begin{aligned} x'(t) &= F(x(t), y(t)) = F(x_0, y_0) \\ y'(t) &= G(x(t), y(t)) = G(x_0, y_0) \end{aligned}$$

Thus, if there is such a constant solution, it must be the case that

$$F(x_0, y_0) = \mathrm{G}(x_0, y_0) = 0.$$

Conversely, if there is a point (x_0, y_0) in the plane at which both $F(x_0, y_0)$ and $G(x_0, y_0)$ equal zero, then certainly the constant functions $x(t) = x_0, y(t) = y_0, -\infty < t < \infty$ form a solution of the system.

The first step in the analysis of an autonomous system of differential equations is to locate these special points.

DEFINITION Any point (x_0, y_0) in the plane at which the functions F and G are both zero is called a *critical point* of the system. Any other point in the plane is called a *regular point*.

Example

The autonomous system

$$\begin{aligned} \frac{dx}{dt} &= x_2 + y_2 - 100 = F(x, y) \\ \frac{dy}{dt} &= x - 2y + 10 = G(x, y) \end{aligned}$$

has two critical points, (–10, 0) and (6, 8). The point (–8, 6) is a regular point, since $G(-8, 6) = -10 \neq 0$, even though $F(-8, 6) = 0$. The point (0, 0) is also a regular point, because $F(0, 0) = -100$ while $\mathrm{G}(0, 0) = 10$.

The critical points for this example can be found by graphing the curves $F(x, y) = 0$ and $G(x, y) = 0$ and determining their points of intersection. Here we have the intersection of a circle and a straight line (see Fig. 4.1).

Other names for critical points are *singular points*, *stable points*, *points of equilibrium*, and *equilibrium states*. You may think of a critical point as a point where the motion described by the pair of differential equations of the system is in a state of rest;

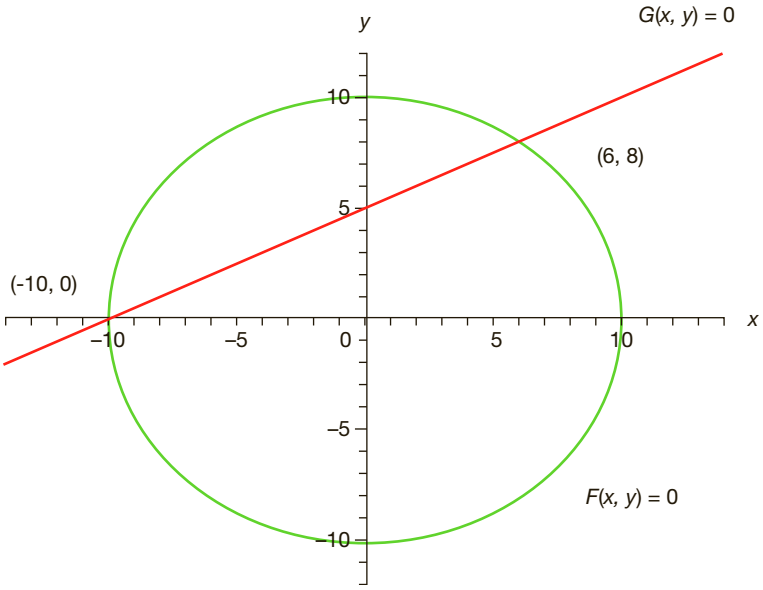


FIGURE 4.1 The stable curves and critical points for the autonomous system $dx/dt = x^2 + y^2 - 100$, $dy/dt = x - 2y + 10$.

both horizontal velocity (dx/dt) and vertical velocity (dy/dt) are zero. At a critical point, both rates of change are zero so that if the orbit starts at such a point, it remains there forever. You can find the critical points by determining the intersections of the two "stable curves" $F(x, y) = 0$ and $G(x, y) = 0$. These curves are also called the *nullclines* of the system.

Simple curve orbits

If an orbit begins at a position that is not a critical point, then at least one of the rates of change, dx/dt or dy/dt, will be nonzero, and the orbit will move away from the point. In an autonomous system, only two things are possible:

1. The orbit will never return to the starting point. This is illustrated in Fig. 2.1 where the orbit is a piece of one branch of a hyperbola.
2. If the orbit ever returns to the starting point, it will simply retrace the same closed curve over and over again. As an example of this, consider again the solution $x(t) = \sin t$, $y(t) = \cos t$ of the system $dx/dt = -y$, $dy/dt = x$ whose orbit is the unit circle.

The orbit for an autonomous system can never cross itself to produce a path—for example, like that traced out by a figure eight (see Fig. 4.2). This is true because the velocities at any point, $(x'(t_1), y'(t_1))$, are completely determined by the coordinates of the point $(x(t_1), y(t_1))$. If we come back to this point at a later time t_2, then the velocities are the same: $x'(t_2) = x'(t_1) = F(x(t_1), y(t_1))$ and $y'(t_2) = y'(t_1) = G(x(t_1), y(t_1))$. In particular, the slopes of the tangent lines to the curve are the same at the two times, and so the direction of motion is exactly the same both times.

In an autonomous system, the orbit is traversed in a fixed direction determined by the system of equations. The direction could only be reversed if a critical point is reached or if

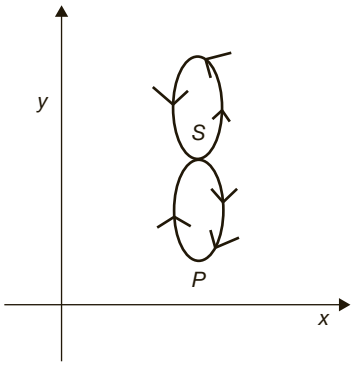


FIGURE 4.2 If the orbit begins at P, then the first time it reaches the point S, the tangent line will have positive slope; the second time it has negative slope. This curve cannot be the orbit of an autonomous system of differential equations.

the curve crosses itself. We have seen that neither of these is possible if the orbit contains a regular point.

Although the orbit can never actually reach a critical point if it does not begin at one, it is possible to approach a critical point asymptotically. As an example, consider the system

$$\begin{aligned} dx/dt &= -x \\ dy/dt &= -y \end{aligned}$$

This system has one critical point, the origin (0, 0). This corresponds to the constant solution $x(t) = y(t) = 0,\ -\infty < t < \infty$. Another solution to the system is the pair $x(t) = y(t) = e^{-t},\ -\infty < t < \infty$. This solution describes parametrically an orbit that is the subset of the line $y = x$ lying in the positive first quadrant. Since $\lim_{t\to\infty} e^{-t} = 0$, the points of this orbit asymptotically approach the origin as time increases.

We can obtain an approximation for the orbit of an autonomous system by using the Euler method (see Chapter 2). For the system

$$\begin{aligned} x'(t) &= F(x, y) & x(0) &= x_0 \\ y'(t) &= G(x, y) \quad \text{with} & y(0) &= y_0 \end{aligned}$$

the sequence $\{P_{i+1}\}$ where $P_0 = (x_0, y_0)$ and $\mathrm{P}_{i+1} = (x_{i+1}, y_{i+1})$ and

$$\begin{aligned} x_{i+1} &= x_i + F(x_i, y_i)\Delta t \\ y_{i+1} &= y_i + G(x_i, y_i)\Delta t \end{aligned}$$

generally provides a good approximation of the exact orbit if Δt is small and the functions F and G are well behaved—for example, if both are differentiable. Other approximation schemes, such as the Runge-Kutta method (see Suggested Project 13), may provide more accurate pictures for the same-sized Δt.

D. Behavior Near a Critical Point

It is of some interest to discover how an orbit behaves in the neighborhood of a critical point. In the ecological models, a critical point corresponds to a "steady state" of zero

population growth or decline for both species. What happens if population levels are near a critical point, but not exactly at it?

We will be looking at three different kinds of behavior that may occur:

1. *Stable equilibrium*: Every orbit near a critical point always approaches it asymptotically.
2. *Unstable equilibrium*: Orbits starting near the critical point always proceed away from it.
3. *Cyclical behavior*: The orbits move around the critical point in tracing out simple closed curves.

Examples of (1) and (2) occurred in the analysis of Richardson's arms race model. Cyclical behavior appears in the system $dx/dt = -y$, $dy/dt = x$, which has the origin as its only critical point. The other orbits are circles centered at the origin. See also Figs. 4.3–4.5.

The basic properties of autonomous systems that have been developed here make possible fruitful analysis of the ecological models presented at the beginning of this chapter.

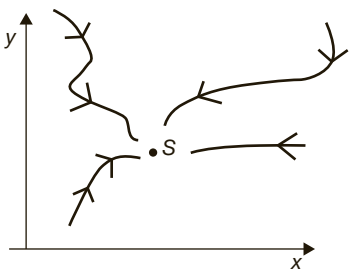


FIGURE 4.3 Stable equilibrium. Orbits that pass through regular points near the critical point S asymptotically approach S.

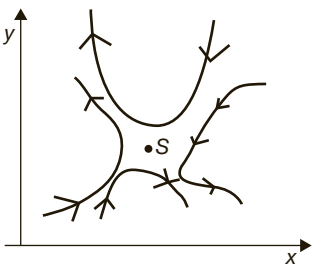


FIGURE 4.4 Unstable equilibrium. Orbits that pass through regular points near the critical point tend to move away from the critical point.

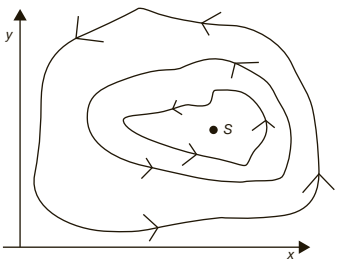


FIGURE 4.5 Cyclical behavior. The orbits passing near the critical point form simple closed curves moving around the critical point.

IV. The Competitive Hunters Model

A. Initial Analysis

As in the study of the Richardson arms race model, analysis begins by locating the critical points, the points where both time derivatives are zero. Since $dx/dt = x(a - by)$, note that $dx/dt = 0$ along the lines $x = 0$ and $y = a/b$. Similarly, $dy/dt = 0$ along the lines $y = 0$ and $x = m/n$, since $dy/dt = y(m - nx)$. There are two critical points: (0, 0) and $(m/n, a/b)$. Each of these single points represents a possible orbit. If initially there are no members of either species, then obviously there can be no gain or loss of any individuals. If there are exactly m/n members of one species and a/b of the other at the start, then the populations will remain at these levels indefinitely, according to the model.

We can also readily identify several other orbits for the competitive hunters system. If $x = 0$ and y is positive at some instant, then at that moment, $dx/dt = 0$ while $dy/dt = my > 0$. The population of the first species will remain at zero while the population of the second is increasing. Geometrically, this means that the positive y-axis is a possible orbit. By a similar argument, the positive x-axis is shown to be an orbit.

The one-point and open ray orbits just found are, of course, quite special and do not indicate the shape of a more typical orbit. They give information, however, that helps determine what those other orbits look like. For example, the fact that orbits cannot intersect each other implies that an orbit that begins in the interior of the first quadrant must always remain there; the boundaries of the first quadrant are made up of other orbits. Thus, if initially there are positive numbers of each species present, then there will always be positive numbers.

Continuing the analysis in the spirit of the Richardson model, note that the lines $y = a/b$ and $x = m/n$ divide the first quadrant into four rectangular regions. The derivative dx/at is positive whenever $y < a/b$ and is negative when $y > a/b$. The derivative dy/dt is positive if $x < m/n$ and negative if $x > m/n$. These facts help establish the general drift of the various orbits. These are indicated in Fig. 4.6.

Fig. 4.6 indicates that if initial population levels are in region IV where $x > m/n$ and $y < a/b$, then the population of the x species will increase, while the population of the y species will decrease. The orbit would remain in region IV.

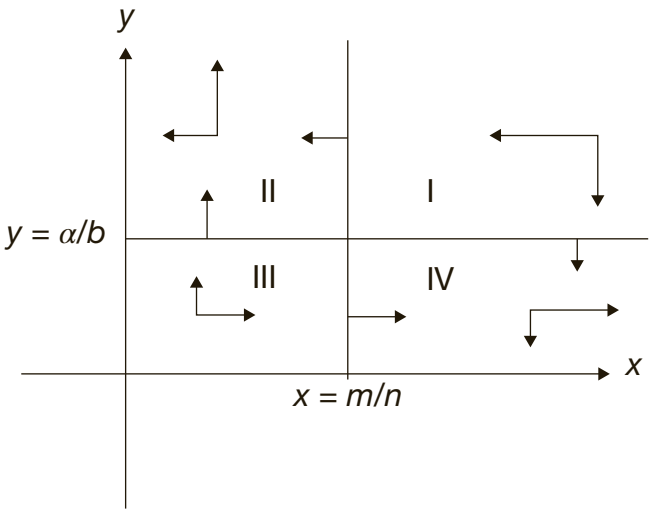
On the other hand, if initially the population of the y species is above its critical level of a/b while the numbers of the x species are below the critical level m/n, then the former species will flourish and the latter will decline. An orbit beginning in region II remains in this region.

If both species are initially below their critical level, the orbit will begin in region III. Analysis of the signs of dx/dt and dy/dt shows that both species will increase in numbers for a while, but the ultimate behavior is unclear. The orbit might enter region IV, enter region II, or asymptotically approach the critical point.

Analogous remarks may be made if the initial populations of both species are above the critical levels, although in that case, both populations will decrease at the start.

B. Further Analysis

There is a powerful technique that sheds further light on the qualitative behavior of an orbit of an autonomous system of differential equations. It is based on a theorem that asserts that the nature of an orbit near a critical point S of the system $dx/dt = F(x, y)$, $dy/dt = G(x, y)$



Region	$\frac{dx}{dt}$	$\frac{dy}{dt}$
I	< 0	< 0
II	< 0	> 0
III	> 0	> 0
IV	> 0	< 0

FIGURE 4.6 Signs of dx/dt and dy/dt for the competitive hunters model.

may be determined by expanding F and G in Taylor series about the point S and retaining only the linear terms. The solutions of these linear equations near the critical point will often have the same general qualitative nature as the exact solutions.

Some explanations are necessary here. Suppose that $z = f(x, y)$ is a function of x and y that behaves nicely near the point $(p, q) = S$. Let h and k be small numbers. A *Taylor series* expansion of f about S is an infinite series of the form

$$f(p,q) + a_1h + a_2k + a_3h^2 + a_4k^2 + a_5hk + a_6h^3 + a_7h^2k + \cdots$$

where each term is a constant multiple of a product of a power of h and a power of k. The coefficients a_i are found by evaluating partial derivatives of f of various orders at the point $S = (p, q)$. The first two coefficients are $a_1 = f_x(p, q)$ and $a_2 = f_y(p, q)$.

If the function f is "nice" near (p, q), then the series converges to the value $f(p + h, q + k)$. More exactly, if there is some circle centered at (p, q) inside of which all the partial derivatives of f of all orders are continuous, then the series converges to $f(p + h, q + k)$ whenever $(p + h, q + k)$ lies inside that circle.

Terminating the Taylor series at a finite number of terms would then give an approximation to the value of $f(p + h, q + k)$. In particular, a crude approximation may be obtained by using only the first three terms—that is,

$$f(p + h, q + k) \sim f(p, q) + f_x(p, q)h + f_y(p, q)k.$$

This approximation is a good one provided that h and k are both small in absolute value; the error, in fact, is bounded by $Ah^2 + Bhk + Ck^2$ for fixed constants A, B, C. In this

approximation, we have neglected all powers of h and k beyond the linear terms. (A fuller treatment of the Taylor series expansion is given in most textbooks on several variable calculus; one reference is Section 3.2 of J. E. Marsden and A. J. Tromba, *Vector Calculus*, New York: W. H. Freeman, 2011.)

In the competitive hunters model, the function $F(x, y)$ has the form $F(x, y) = ax - bxy$ so that $F_x(x, y) = a - by$ while $F_y(x, y) = -bx$. At the critical point $S = (m/n, a/b)$, both functions F and F_x are zero, while $F_y(m/n, a/b) = -bm/n$. Applying the linearized Taylor series

$$F\left(\frac{m}{n} + h, \frac{a}{b} + k\right) \approx \frac{-bm}{n}k$$

approximation with $f = F$, $p = m/n$, and $q = a/b$, we conclude that

$$F\left(\frac{m}{n} + h, \frac{a}{b} + k\right) \approx \frac{-bm}{n}k$$

A similar analysis for $G(x,y) = my - nxy$ shows that

$$G\left(\frac{m}{n} + h, \frac{a}{b} + k\right) \approx \frac{-an}{b}h$$

Define, next, two new variables u and v by $u = x - (m/n)$ and $v = y - (a/b)$. Then $\frac{du}{dt} = \frac{dx}{dt}$ and $\frac{dv}{dt} = \frac{dy}{dt}$. Furthermore, $x = (m/n) + u$ and $y = (a/b) + v$. The Taylor series approximation can be rewritten as

$$F(x,y) \approx -\frac{bm}{n}v$$

$$G(x,y) \approx -\frac{an}{b}u$$

Thus, we have

$$\frac{du}{dt} = \frac{dx}{dt} = F(x,y) \approx -\frac{bm}{n}v$$

$$\frac{dv}{dt} = \frac{dy}{dt} = G(x,y) \approx -\frac{an}{b}u$$

By the theorem on the general nature of the orbit, we conclude that the orbits near the critical point of the original system behave like the orbits of the simpler system

$$\frac{du}{dt} = -\frac{bm}{n}v$$

$$\frac{dv}{dt} = -\frac{an}{b}u$$

The orbits of this simpler system are obtained by first noting that the chain rule gives

$$\frac{du}{dv} = \left(\frac{b^2 m}{an^2}\right)\frac{v}{u}$$

and separation of variables yields

$$\int an^2 u\,du = \int b^2 m v\,dv$$

Integrate and rewrite to obtain

$$an^2 u^2 - b^2 m v^2 = K$$

where K is an integration constant.

Rewrite this last equation in terms of the original variables

$$am^2\left(x - \frac{m}{n}\right)^2 - b^2 m\left(y - \frac{a}{b}\right)^2 = K$$

which is the equation of a hyperbola in the xy-plane with center at $(m/n, a/b)$. The value of the constant K depends on the initial population levels of the two species. Once these are known, K may be determined, and with it which of the two branches of the hyperbola represents the actual orbit. One branch of the hyperbola asymptotically approaches the x-axis, and the other asymptotically approaches the y-axis.

The qualitative behavior of the orbits of the original competitive hunters model that pass close to the critical point must be like the qualitative behavior of these hyperbolas—that is, either the x values increase indefinitely as the y values tend to zero or the y values increase indefinitely as the x values tend to zero.

C. Exact Orbits

To obtain the exact orbits for the competitive hunters model, note that the equations of the system

$$\frac{dx}{dt} = ax - bxy, \quad \frac{dy}{dt} = my - nxy$$

may be combined into a single first-order differential equation:

$$\frac{dy}{dx} = \frac{my - nxy}{ax - bxy} = \frac{y(m - nx)}{x(a - by)}$$

Separating the variables and integrating gives

$$\int \frac{a - by}{y}dy = \int \frac{m - nx}{x}dx$$

and, when the aniderivatives are found,

$$a \log y - by = m \log x - nx + C$$

where C is an integration constant.

Exponentiate each side to obtain

$$y^a e^{-by} = Kx^m e^{-nx}$$

where K is the constant e^C.

It is not possible to solve this last equation to obtain y as an explicit function of x. However, it is possible, thanks to a technique invented by Vito Volterra, to obtain a graph of this relationship in the xy-plane.

Volterra began by noticing that he could graph the functions $v = x^m e^{-nx}$ and $u = y^a e^{-by}$ in the (x, v) and (y, u) planes, respectively, and that these graphs are similar in form. Fig. 4.7 shows this curve for a particular choice of constants m and n.

The initial analysis showed that the orbit will remain in the first quadrant of the (x, y)-plane. The other three quadrants will be used to represent the first quadrants of the (y, u)-, (u, v)-, and (v, x)-planes, respectively. Fig. 4.8 indicates how to do this.

To find a point on the orbit of the solution to the system, use Volterra's procedure:

1. Select a positive value for x—say, x_0.
2. Determine the value v_0 corresponding to x_0 from the equation $v = x^m e^{-nx}$.
3. Determine the value u_0 corresponding to v_0 from the relationship $u = Kv$.
4. Determine the y values (in general, there will be two) corresponding to u_0 by finding where the vertical line through (u_0, v_0) intersects the curve $u = y^a e^{-by}$.
5. Extend horizontal lines through these y values until they intersect the vertical line through $(x_0, 0)$ in the (x, y)-quadrant. These intersections determine points on the orbit.

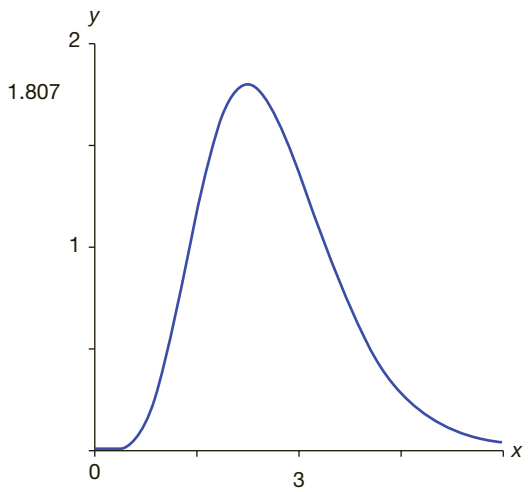


FIGURE 4.7 The graph of $v = x^m e^{-nx}$ with $m = 6$, $n = 2$.

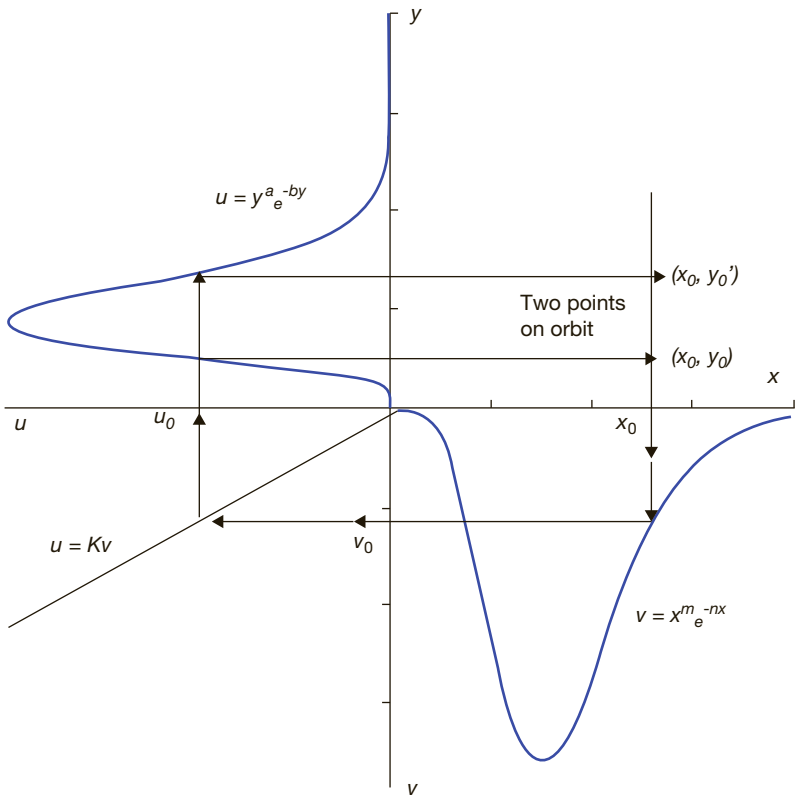


FIGURE 4.8 Illustration of the Volterra mapping technique.

If this procedure is followed for a large number of choices for x_0, an accurate picture of the orbit in the (x, y)-plane emerges (see Fig. 4.9). Note again that each orbit asymptotically approaches one of the coordinate axes, even if the initial point of the orbit is chosen to be relatively far from the critical point. Although the Volterra mapping technique requires careful graphing, it has greater applicability than the analytic technique of using the linearized Taylor series expansion.

D. Interpretation of Results

The mathematical analysis of the competitive hunters model yields several conclusions:

Equilibrium is possible. There is a critical positive population for each species. If each maintains that level, they can coexist in the same environment.

The equilibrium is highly unstable. If, at any instant, the population levels are not at the critical sizes, then the effects of competition will be for one species to flourish and the other to die out. If one species exceeds its critical size while the other fails to achieve its, then the first one emerges triumphant. If both species are either above or below the critical sizes, then more detailed knowledge of the size of the parameters a, b, m, and n and exact numbers of initial population levels must be known to predict which species will win out.

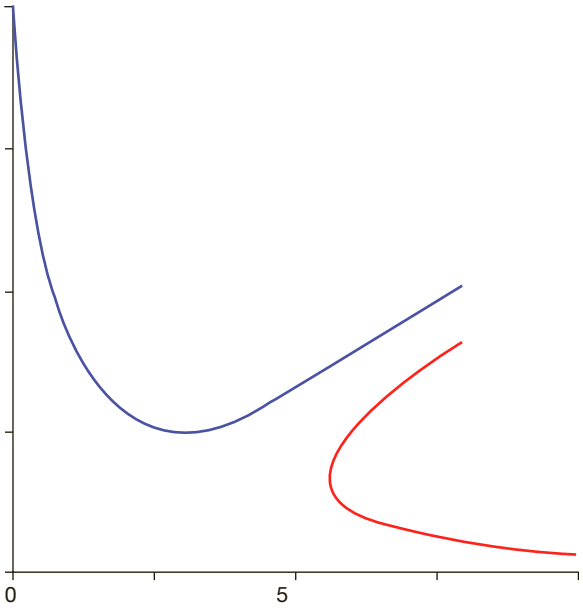


FIGURE 4.9 Orbits of the competitive hunters model. Here $a = .5$, $b = .3$, $m = .6$, $n = .2$. The critical points are (0, 0) and (3, 5/3).

The prediction emerging from the model that only one species is likely to survive is consistent with known biological laws. It is called the *principle of competitive exclusion*. The noted zoologist Ernst Mayr comments in his book *Populations, Species and Evolution* [1970]:

> *The result of competition between two ecologically similar species in the same locality is either (1) the two species are so similar in their needs and their ability to fulfill these needs that one of the two species becomes extinct, either (1a) because it is "competitively inferior," that is, it has a smaller capacity to increase or (1b) because even though competitively equivalent it had an initial numerical disadvantage; or (2) there is a sufficiently large zone of ecological nonoverlap (area of reduced or absent competition) to permit the two species to coexist indefinitely. In sum: two species cannot indefinitely coexist in the same locality if they have identical ecological requirements. This theorem is sometimes referred to as the Gause principle, after the Russian biologist Gause who was the first to substantiate it experimentally. Yet . . . the principle was known long before Gause. Darwin discussed it at length in his* Origin of Species. . . . *The validity of this exclusion principle has been tested in numerous laboratory experiments in which mixed populations of two species were established in a uniform environment. In virtually every case, one of the two species was eliminated sooner or later.*

How accurately the principle of competitive exclusion describes what occurs in the very complex ecosystems that actually characterize the real natural world is a source of much discussion among ecologists. As Thomas Ray notes:

> *Although the principle of competitive exclusion has been experimentally demonstrated in the laboratory, and is considered theoretically sound, natural communities widely flout the principle. In tropical rain forests, for example, several hundred species of trees coexist without any dominant species in the community. All species of trees must spread their leaves to collect light*

and their roots to absorb water and nutrients. Evidently there are not several hundred niches for trees in the same habitat. Somehow the principle of competitive exclusion is circumvented.

There are many theories on how competitive exclusion may be circumvented. One leading theory is that periodic disturbance at the proper level sets back the process of competitive exclusion, allowing more species to coexist. There is substantial evidence that moderate levels of disturbance can increase diversity.

E. Modifying the Model

The competitive hunters model makes two major predictions:

1. One species will die out.
2. The other species will grow indefinitely numerous.

The first prediction, as just noted, is consistent with many observations and experiments. The second is not. The source of this second prediction is one of the assumptions made in building the model: in the absence of one of the species of hunters, the other species increases at a rate proportional to its population size—that is, it would experience exponential growth.

To improve the model, this assumption should be replaced by a more realistic one. Perhaps the assumption should be that in the absence of one species, the other species experiences logistic growth. The reader is invited to formulate a model built on this assumption and to derive the appropriate mathematical conclusions and real-world interpretations from it.

V. The Predator-Prey Model

A. Analysis

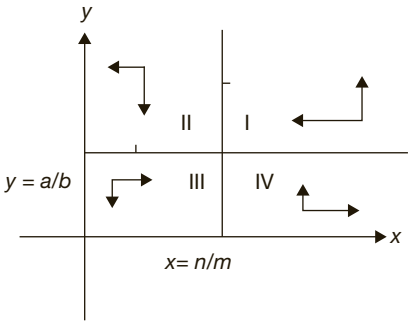
We turn now to an examination of the predator-prey model, the system

$$\frac{dx}{dt} = ax - bxy = x(a - by)$$

$$\frac{dy}{dt} = mxy - ny = y(mx - n)$$

where a, b, m, n are positive constants, x is the population of prey (gazelles), and y is the population of predators (leopards). As in the case of the competitive hunters model, $dx/dt = 0$ along the lines $x = 0$ and $y = a/b$, while $dy/dt = 0$ on the lines $y = 0$ and $x = n/m$. The critical points are (0, 0) and $(n/m, a/b)$. The positive x-axis and the positive y-axis are also orbits. All other orbits of interest are contained entirely in the first quadrant of the xy-plane.

The lines $y = a/b$ and $x = n/m$ divide the first quadrant into four rectangular regions. The differences between the predator-prey model and the competitive hunters model become evident when the signs of the derivatives dx/dt and dy/dt are determined in each of these four regions. See Fig. 4.10. Note that dx/dt is positive if $y < a/b$ and negative if $y > a/b$, while dy/dt is positive whenever $x > n/m$, but negative if $x < n/m$. The general drift of the orbits of the system is evident from Fig. 4.10.



Region	$\frac{dx}{dt}$	$\frac{dy}{dt}$
I	< 0	> 0
II	< 0	< 0
III	> 0	< 0
IV	> 0	> 0

FIGURE 4.10 Signs of dx/dt and dy/dt for the predator-prey model.

No matter where the initial population levels are located, the orbit will follow a counterclockwise direction about the critical point. For example, if there are a small number of gazelles and leopards at the start (initial level in region III), then the gazelle population will increase at first, while the leopard population decreases. The small number of leopards poses little threat to the gazelles, while the scarcity of gazelles means that the leopards have a difficult time finding ample food.

When the gazelle population reaches a critical level of n/m, then the leopard population also begins to increase. For a time, while the orbit is in region IV, both species experience a growth in numbers. Eventually the leopard population exceeds its critical level of a/b. Now the leopards are sufficiently plentiful to endanger the gazelle population, whose numbers begin to decline while the leopard population increases; the orbit is in region I. When the gazelle population declines below n/m, as the orbit enters region II, then there is not a sufficient supply of prey to sustain a large leopard population. Both species lose numbers until the orbit reaches region IV again.

The fluctuations of the populations then seem to be following a cyclical pattern of some sort. What is not clear from this initial analysis is whether the orbits are spiraling toward the critical point, spiraling away from it, or possibly exhibiting some other type of oscillation. To answer this question, consider the linearized Taylor series expansion.

The functions to be approximated are $F(x, y) = ax - bxy$ and $G(x, y) = mxy - ny$. The calculations yield

$$F\left(\frac{n}{m} + h, \frac{a}{b} + k\right) \approx -\frac{bn}{m}k$$

and

$$G\left(\frac{n}{m} + h, \frac{a}{b} + k\right) \approx \frac{am}{b}h$$

Make the change of variables $u = x - (n/m)$, $v = y - (a/b)$ so that

$$\frac{du}{dt} = \frac{dx}{dt} = F(x,y) \approx -\frac{bn}{m}v$$

$$\frac{dy}{dt} = \frac{dv}{dt} = G(x,y) = \frac{am}{b}u$$

The orbits near the critical point of the predator-prey system will have the same general behavior as the orbits of the simpler system

$$\frac{du}{dt} = -\frac{bn}{m}v$$

$$\frac{dy}{dt} = \frac{am}{b}u$$

(For a proof of this claim, see Chapter 8 of Bruce P. Conrad, *Differential Equations*: *A Systems Approach*, Upper Saddle River, NJ: Prentice Hall, 2003.) Using the fact that

$$du/dv = -\left(b^2n/am^2\right)(u/v)$$

we separate variables, integrate, and conclude that the simpler system has a solution satisfying

$$am^2u^2 + b^2nv^2 = K$$

where K is a constant of integration. Rewriting this equation in terms of the original variables gives

$$am^2\left(x - \frac{n}{m}\right)^2 + b^2n\left(y - \frac{a}{b}\right)^2 = K$$

This is the equation of an ellipse with center at $(n/m, a/b)$ and with axes parallel to the coordinate axes of the xy-plane. Near the critical point, the orbits are elliptical trajectories centered at the critical point. The orbits do not spiral toward the point or away from it (see Fig. 4.11).

It is possible to solve the simpler system for u and v explicitly as functions of t. This is done by computing second derivatives with respect to t:

$$u'' = \frac{-bn}{m}v' = \frac{-bn}{m}\frac{am}{b}u = -anu$$

$$v'' = \frac{am}{b}u' = \frac{am}{b}\frac{-bn}{m}v = -anv$$

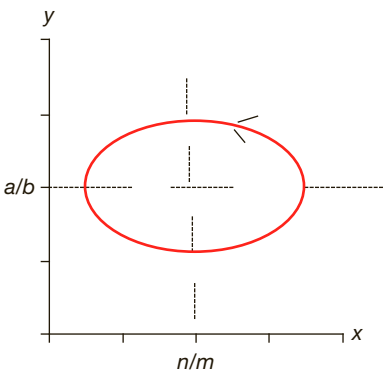


FIGURE 4.11 Elliptical orbit of the linearized version of the predator-prey model.

Note that both these equations are of the form $z'' = -pz$, where p is a positive constant. The general solution of such a second-order differential equation is $z = A\sin\sqrt{p}t + B\cos\sqrt{p}t$, where A and B are constants. Thus, z is a periodic function with period $2\pi/\sqrt{p}$.

The solution of the simpler system is a pair of functions of t, u, and v, with the same period $2\pi/\sqrt{p}$. Recalling that the average value of a continuous function f on an interval [a,b] is defined as $\dfrac{\int_\alpha^\beta f(t)dt}{\beta-\alpha}$, it is easy to check whether u and v have average values of 0. Since $u = x - (n/m)$ and $v = y - (a/b)$, this means that x and y would have average values of n/m and a/b, respectively. The conclusion is that near the critical point the trajectories display periodic movement and are approximated by ellipses with period $2\pi/\sqrt{p}$.

The Volterra mapping technique can be used to find a more exact orbit to the original predator-prey model. Note that

$$\frac{dy}{dx} = \frac{G(x,y)}{F(x,y)} = \frac{y(mx-n)}{x(a-by)}$$

in the original system. After the variables are separated in this differential equation and integration is completed, the solution looks like

$$a\log y - by = mx - n\log x + C$$

which we can rewrite as

$$\left(y^a e^{-by}\right)\left(x^n e^{-mx}\right) = K$$

For any particular choice of constants a, b, m, n, K, Volterra's method gives a graph of the set of all points (x, y) satisfying the equation. The only modification required in the procedure of Section IV.C is in Step 3, where the relationship $uv = K$ must now be used in place of $u = Kv$. Note that Volterra's method shows that for each x-value, there are at most two y-values. Thus, the orbit for the predator-prey can not be spiral, for in a spiral, there would be some vertical line (a particular x-value) that hit the orbit infinitely many times. Fig. 4.12 shows a typical result.

B. Interpretation and Testing of Results

Alfred J. Lotka was the first person to formulate and study closely mathematical models of interacting populations. In his 1925 book *Elements of Physical Biology*, Lotka considered a wide variety of relationships that can occur between two species, including the models presented in this chapter. Vito Volterra began to consider such models at the request of a zoologist, Umberto D'Ancona, who was studying the variations in numbers of fishes caught in the Adriatic during the period of World War I. Beginning in 1926, Volterra developed a mathematical analysis for interactions among any number of species.

The early work of Lotka and Volterra has been revised and generalized by many mathematicians and mathematical biologists and many experiments have tested the conclusions of their models in laboratory situations. The simple predator-prey model devised by Lotka and Volterra predicts oscillations in the numbers of the two species. Such

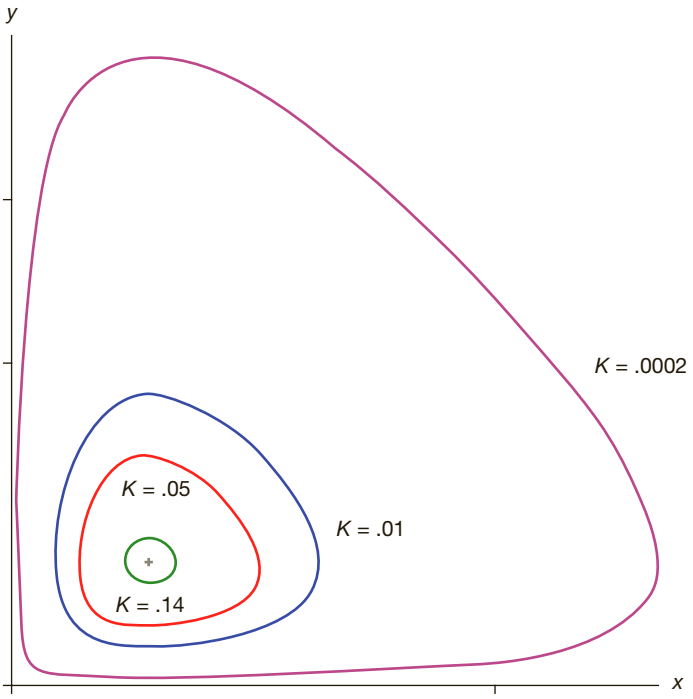


FIGURE 4.12 Orbits for the predator-prey system $dx/dt = ax - bxy$, $dy/dt = mxy - ny$ obtained from the Volterra mapping technique. Here $a = 5$, $b = 3$, $m = 2$, and $n = 6$. Critical points are (0, 0) and (3, 5/3). All orbits move counterclockwise. The constant K is determined by initial conditions (x_0, y_0) and is equal to $\frac{x_0^n y_0^a}{e^{by_0 + mx_0}}$. At (3, 5/3), $K = .156$.

oscillations have been observed in experiments, but only in fairly complex ones. A common outcome of simpler (and thus less natural) experiments is that the predators devour all the prey and then die out themselves. In a carefully designed experiment, C. B. Huffaker in 1957 created a predator-prey oscillation using as prey a mite that feeds on oranges and another species of mite as its predator. The Lotka-Volterra model compares reasonably well with the observed data.

Population oscillations in the world have also been observed. E. R. Leigh, in a 1969 study, concluded that the fluctuations in the numbers of Canadian lynx and its primary food, the hare, trapped by the Hudson's Bay Company between 1847 and 1903 were periodic. The observed period is not in good agreement with that predicted by the Lotka-Volterra model. This may be because the numbers of animals trapped were not a fair representative sampling of the actual populations, but more likely there are other environmental factors affecting the lynx and the hare that are not included in the model.

An interesting property of the predator-prey model is revealed by considering the effect of removing both species from the community in quantities proportional to their numbers. This commonly happens when the environment is subject to pesticide sprays inimical to both species. The effect is reflected by a decrease in the coefficient a and an increase in the coefficient n in the differential equations defining the model. Since the average number of predators is about a/b and that of prey is about n/m, the long-term consequences are to decrease the average predator population while *increasing* the average number of prey. One moral is clear: it can be self-defeating for man to use an insecticide against a species whose population is already being controlled by a natural predator.

An example reinforces this observation. The accidental introduction in the United States of the cottony cushion insect *icerya purchasi* from Australia in 1868 threatened to destroy the American citrus industry. To counteract this, a natural Australian predator, a ladybird beetle (*novius cardinalis*) was imported. The beetles kept the scale insects down to a relatively low level. When DDT was discovered to kill scale insects, farmers applied it in the hopes of reducing further the scale insect population. DDT, however, was also fatal to the beetle; the overall effect of using insecticide was to increase the numbers of the scale insect.

C. Modifying the Model

Several variations of the Lotka-Volterra predator-prey model have been proposed that offer more realistic descriptions of the interactions of the populations.

1. If the population of gazelles is always much larger than the number of leopards, then the considerations that entered into the development of the logistic equation may come into play. If the number of gazelles becomes sufficiently great, then the gazelles may be interfering with each other in their quest for food and space. One way to describe this effect mathematically is to replace the original model with the more complicated system

$$\frac{dx}{dt} = ax - bx^2 - cxy$$

$$\frac{dy}{dt} = mxy - ny$$

where a, b, c, m, n are positive constants.

2. Most predators feed on more than one type of food. If the leopards can survive on an alternative resource, although the presence of their natural prey (gazelles) favors growth, a possible alternative model is the system

$$\frac{dx}{dt} = ax - bx^2 - cxy$$

$$\frac{dy}{dt} = mxy + ny - py_2$$

where a, b, c, m, n, p are positive constants.

3. P. H. Leslie and J. C. Gower studied a third variation, the system of equations

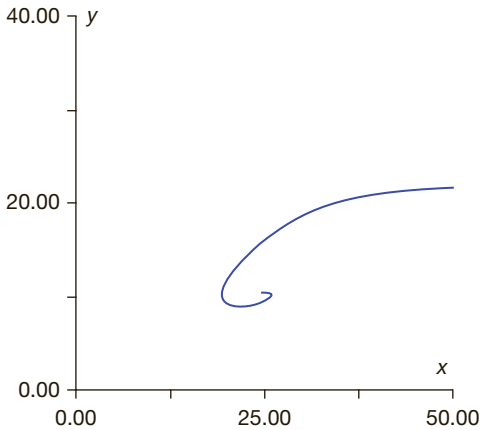
$$\frac{dx}{dt} = ax - bxy$$

$$\frac{dy}{dt} = \left(c - e\frac{y}{x}\right)y$$

where the parameters a, b, c, e are again positive constants. Here the term y/x arises from the fact that this ratio ought to affect the growth of the predator. When leopards are numerous and gazelles are scarce, y/x is large and the growth of leopard population will be small. Conversely, when the supply of gazelles is ample for the leopards, y/x is small and there is slight restriction on the increase of the predators.

The orbits associated with the Leslie-Gower model are curves that spiral in toward the critical point. Fig. 4.13 illustrates a typical situation of this stable equilibrium.

4. The original predator-prey model and the variations just discussed all reflect an assumption that the predators are insatiable: there is no upper limit to the amount of prey they will consume. In reality, however, limits on gut size and time available for hunting indicate that the consumption rate should approach an upper bound as the prey density increases.



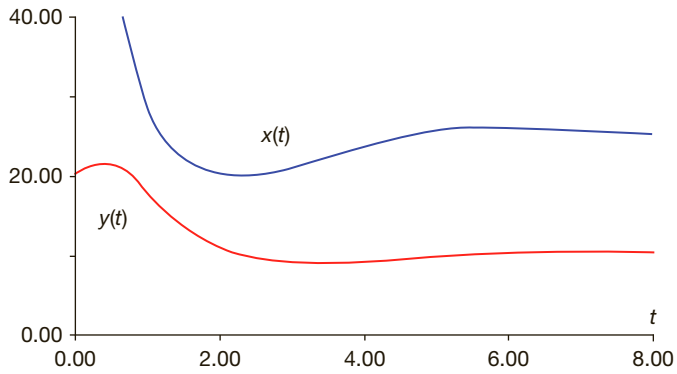


FIGURE 4.13 Results of the Leslie-Gower model for a predator-prey system. Here $dx/dt = ax - bxy$, $dy/dt = (c - e(y/x))y$. The curves illustrated are for $a = 1$, $b = .1$, $c = 1$, $e = 2.5$ and initial populations $x_0 = 80$, $y_0 = 20$. The critical point is (25, 10). The top graph shows the orbit of a solution of the system of differential equations; it spirals in toward the critical point. The bottom graph shows x and y as functions of t. From E. Pielou, *An Introduction to Mathematical Ecology*, New York: John Wiley, 1969. Reprinted by permission of the publisher and author.

Gary Harrison [1995] proposed a generalization of the Lotka-Volterra equations, which includes a term, $f(x)$, measuring a *functional response* of the predator as prey population increases. His equations have the form

$$\frac{dx}{dt} = ax\left(1 - \frac{x}{k}\right) - bf(x)$$

$$\frac{dy}{dt} = mf(x)y - ny$$

$$\frac{dx}{dt} = ax\left(1 - \frac{x}{k}\right) - bf(x)y$$

$$\frac{dy}{dt} = mf(x)y - ny$$

Note that Harrison's approach also uses the more realistic assumption that prey experience logistic growth in the absence of predators. Michael L. Rosenzweig and Robert H. MacArthur studied this model using the function

$$f(x) = \frac{x}{c + x}$$

which gives a "saturation" effect in functional response when prey is abundant. Fig. 4.14 shows two typical orbits for the Rosenzweig-MacArthur model. Predator and prey populations do oscillate over time, but an orbit does not close up; it approaches a closed curve, called a *stable limit cycle*.

The detailed development and analyses of these variations and the creation of new ones are left as suggested modeling projects for the reader.

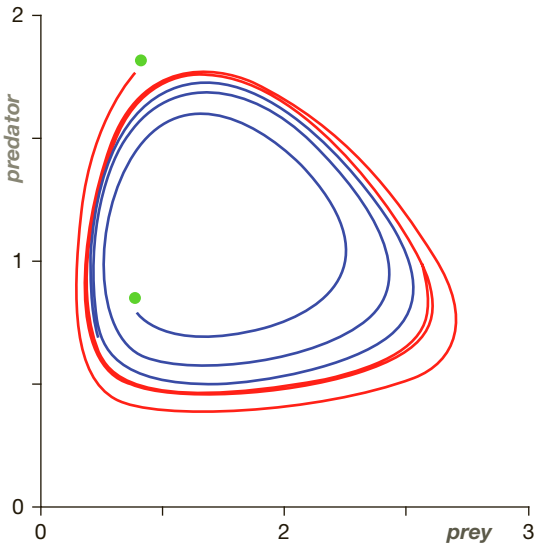


FIGURE 4.14 Two trajectories in the (prey, predator)-plane converging to a stable limit cycle in Harrison's model.

VI. Concluding Remarks on Simple Models in Population Dynamics

"Criticizing mathematical models in ecology," the American mathematical biologist George Oster once wrote, "is like harpooning a blimp; it is almost impossible to miss, and every thrust is likely to be fatal!"

It is easy to list many ecological factors that have been omitted from the simple models considered here:

1. Nonuniformity of the environmental conditions. The ecological system under investigation will not be uniform in either space or in time. The simple models will then have their best validity only over small geographical areas and short periods of time.
2. Individual differences in organisms constituting the population. The growth rate for a population of gazelles, for example, is only an average for the entire population and may differ markedly among individuals, especially those of differing ages.
3. Immigration and emigration. Except in carefully controlled laboratory experiments, the ecosystem is not isolated from the rest of the world. Animals may enter or leave at any time.
4. Spatial clumping of the organisms so that the effects of density dependence will not be the same everywhere.
5. Effects of time lag in the response of organisms to environmental change. The simple differential equations models assert that growth rates adjust instantaneously to changes in population levels.
6. Effects of other species that interact with the system. Gazelles have other enemies than leopards, for example, while leopards do not limit their diet to gazelles.
7. Random disturbances. An unexpected fire, flood, or epidemic affects population levels immediately and often with catastrophic results.

Even though we have not taken into account these and many other factors, it sometimes is found that actual populations behave in a manner very similar to that predicted by the simple models. There are several possible explanations for this:

- The factors neglected may indeed be of negligible importance.
- Some of the neglected factors may be important, but may cancel each other out.
- The resemblance of a model to the real-life process it is intended to represent may not be as close as it seems. Closer investigation of the predictions of the model and the actual situation may reveal crucial differences.

These three explanations should always be considered in the evaluation of any mathematical model of a real-world phenomenon.

Scholars in fields far removed from ecology have found the models we have studied in this chapter to be useful in the analysis of other situations that involve human interactions. We'll briefly mention a few here.

In trying to gain a deeper understanding of revolutions, Jason Epstein initiated the use of a variation of a predator-prey model. Noting that revolutionaries have widely varying

goals—overthrow a monarchy, install a theocracy, establish a democracy—Epstein sought a mathematical model whose dynamic behavior mimicked revolutionary processes in general, "regardless of their political 'substance.'"

Epstein considers a population divided into two groups whose populations change over time. Here $y(t)$ represents the number of individuals actively involved in articulating "a revolutionary vision" and winning over converts from a population $x(t)$ of persons thought to be receptive the idea of revolution. The revolutionaries gain in numbers by winning over individuals to their side but lose strength when they are imprisoned by the authorities. One mathematical model Epstein considers has the form

$$\frac{dx}{dt} = ax - bxy$$

$$\frac{dy}{dt} = bxy - cy$$

where the positive constants a, b, and c measure, respectively, the intrinsic growth rate of the x population, the conversion rate, and the imprisonment rate.

Richard Goodwin employed a Lotka-Volterra model to study cycles in economic growth rates. Goodwin's model incorporates many factors such as output, capital, wage rates, labor productivity, workers' share of product, labor supply and employment, but ends up as a classic predator-prey pair of differential equations. Roberto Veneziania and Simon Mohunb [2006] describe Goodwin's model as "one of the first and most elegant dynamic formalisations of Marx's theory of distributive conflict and a seminal contribution in the use of non-linear models drawn from mathematical biology to analyse economic phenomena."

Several tongue-in-cheek papers about vampires have been published in serious academic research papers. Most begin with the model [Hartl et al. 1992]:

$$H'(t) = nH - dVH$$
$$V'(t) = dVH - aV$$

where H is the stock of humans in an isolated Transylvanian community and V is the number of vampires. The parameters d, n, and a are positive constants, where d is a contact coefficient, n is the growth rate of the human population, and a is the "death rate of vampires due to contact with sunlight, crucifixes, garlic, and vampire hunters." Note that this pair of equations is equivalent to the Lotka-Volterra model.

Anthropologist Jeffrey Brantingham and his colleagues at UCLA used Lotka-Volterra equations to study the territories of rival gangs. Their model [Brantingham et al. 2012] predicted that 59% of gang crimes would occur within two blocks of a border between two gangs and 87.5% would occur within about three blocks. When the researchers mapped more than 500 crimes attributed to 13 gangs in Los Angeles, they found that, in fact, 58% and 83% occurred within two blocks and three blocks of a border, respectively. This research may eventually be used to identify zones to be more intensively patrolled by police with the goal of disrupting assaults and murders perpetrated by gangs.

In your earlier study of mathematics (see also Chapters 1 and 3), you saw many different applications where the underlying mathematical model was the differential equation for exponential growth. It should not be surprising, then, that the same system of

nonlinear differential equations may arise in the study of a wide variety of situations that, on the surface, may appear to have nothing to do with each other. The term *dynamical analogies* is used to describe processes whose mathematical descriptions have the same functional form where we can put into one-to-one correspondence the variables and parameters. "It is a startling fact," Epstein [1997] notes, "that a huge variety of seemingly unrelated processes are analogous in this sense. . . . Analogy . . . has played a powerful role in the development of science, engineering and also social science."

Employing the same mathematical model to study diverse phenomena is often cited as an example of the unifying power of mathematics and certainly has a strong aesthetic appeal. But it also has a powerful practical purpose. As Harry Olson [1958] argued:

> *Analogies are useful for analysis in unexplored fields. By means of analogies an unfamiliar system may be compared with one that is better known. The relations and actions are more easily visualized, the mathematics more readily applied and the analytical solutions more readily obtained in the familiar system.*

VII. Biographical Sketches

Photograph reproduced by permission of Metropolitan Life Insurance Company

Alfred J. Lotka

A. Alfred James Lotka

The father of demographic analysis, Alfred James Lotka (1880–1949) made many important early contributions to the development of a mathematical approach to the study of social phenomena. Besides his own considerable research on population theory, evolutionary processes, and self-renewing aggregates, Lotka wrote books and articles informing social scientists and the general public of new developments in science and suggesting ways that mathematics might be used to study behavior.

Lotka was born in Lemberg, Austria (formerly Lwów, Poland, and now Lviv, Ukraine), on March 2, 1880. His father Jacob, a convert from Judaism to Christianity,

headed a group of missionaries associated with the London Society for Promoting Christianity Amongst the Jews. Alfred Lotka received his early education in France, but obtained his professional training in England (Birmingham University), Germany (University of Leipzig), and the United States (Cornell, Johns Hopkins). This variety of educational background produced in his works, according to one critic, "a happy alliance of the deductive turn of the French spirit, the pragmatic tendency of the English character, and the Germanic concern for precision and erudition."

After his arrival in the United States in 1902, Lotka worked as a chemist for a commercial chemical company, an assistant in physics at a major university, an editor for *Scientific American* publications, and an examiner for the U.S. Patent Office. In 1924, he joined the statistical bureau of the Metropolitan Life Insurance Company in New York City. During the quarter century he worked for Metropolitan, Lotka developed systematically and in collaboration with others the demographic analysis he had initiated as a young man.

His 95 technical papers and six books comprise "permanent contributions of high scholarly standing" according to Frank W. Notestein. He wrote, "To Dr. Lotka's work, the field of demography owes virtually its entire central core of analytical development." Among his major discoveries was a demonstration of how a closed population (no immigration or emigration) develops a stable age distribution and a characteristic rate of increase. Lotka showed how the intrinsic growth rate should be computed and revealed how misleading is the more naive approach that uses only the crude difference between birth and death rates.

Lotka's most significant impact on the progress of mathematical modeling has been through his book *Elements of Physical Biology*. Originally published more than 80 years ago, it was reissued in 1956 under the more descriptive title *Elements of Mathematical Biology*. In reviewing the book, Herbert A. Simon [1959] discussed its contribution:

> *A sect—and by any reasonable definition, mathematical social scientists formed one—needs arcana, as source both of its special wisdom and of passwords by which its members can recognize each other. In the Thirties, a person who had read Lotka's* Elements of Physical Biology *and Richardson's* Generalized Foreign Policy*, and who was acquainted with the peculiar empirical regularities compiled by Zipf was almost certainly a fellow-sectarian. These works represented a large fraction of the literature, outside of economics, in mathematical social science. . . . It is easy to show that much that has happened in mathematical social science in the thirty years since the publication of the first edition of* Elements of Physical Biology *lies in the direction along which the book points.*

Simon [1959] describes Lotka as a "forerunner whose imagination creates plans of exploration that he can only partly execute, but who exerts great influence on the work of his successors—posing for them the crucial questions they must answer, and disclosing more or less clearly the directions in which the answers lie."

Lotka was widely respected by his colleagues who elected him to the presidencies of the American Statistical Association and the Population Association of America. "Dr. Lotka was a scientist of the first rank, but he was much more," wrote Notestein [1950]. "His popular writings . . . reveal a delicate sense of humor and a deep consideration of the arts. A quiet, learned, modest, and gently humorous man, a wise counselor . . . Dr. Lotka

will always be held in highest esteem by his colleagues of the demographic profession among whom his is the greatest name, and by his friends, who valued the man above his knowledge."

Lotka's attitude toward the role of scientific models can be seen in some excerpts of an article on Einstein's theory of relativity that he wrote for a general audience in 1920 [Lotka, 1920]:

> *One of the foremost aims of science is to build up a conception of the world which shall correspond more and more closely with our experience. As the scope of our experience, our observation, enlarges we shall naturally be forced, from time to time, to modify the world-picture we have already formed. . . .*
>
> *We must seek to overcome mental inertia, to liberate ourselves from preconceived ideas. History has taught us that men are apt to fail to distinguish the* absurd*, the illogical, from the merely* unfamiliar*. Profiting by former experience of the race, we may reasonably expect to cut short our term of apprenticeship. . . .*
>
> *We are so constituted that of the world in which we live we perceive at any instant only one aspect, a snapshot, as it were, taken from the point of space and time at which we happen to be stationed. . . .*
>
> *The thing of paramount importance to us humans, living in a real world, is not what relations* ought to exist *among our observations, but what relations* actually do exist*. If there is disagreement, we shall do well to change our conceptions to fit the facts, for facts are stubborn things which refuse to adapt themselves to fit our conceptions*
>
> *It is not for us to shape the external world in accordance with our concepts; we must build up our conceptual world-picture in accordance with observation.* If a new observation cannot by any manner of means be made to fit into our conception of the world*,* we may be forced to change that conception*.*

B. Vito Volterra

Born in Ancona, Italy on May 3, 1860, Samuel Giuseppe Vito Volterra was an only child of a Jewish family of modest means. His father died when Volterra was barely 2 years old. He began the serious study of arithmetic and geometry at age 11 and was pursuing calculus by the time he was 14. Resisting his family's wishes that he enter a commercial profession, Volterra opted for a scientific career. He was awarded his doctorate in physics from the University of Pisa in 1882. He served first as a professor of mechanics and of mathematics at Pisa and later spent 30 years on the faculty of the University of Rome.

Volterra was the leading Italian mathematician of his day and was a central figure in international academic, intellectual, and political circles. He served as president of the world's oldest scientific society, the Academia dei Lincei (a post once held by Galileo), and founded and headed Italy's National Research Council. When he was 45, Volterra was appointed a senator of the Kingdom of Italy, serving as a member for life of Parliament's upper house.

Volterra's major contributions to pure mathematics lay in the development of functional analysis and the theory of integral equations. By means of the functional calculus, Volterra was able to show that the Hamilton-Jacobi theory of the integration of the differential equations of dynamics could be extended to more general problems of

mathematical physics. His research work on problems of elasticity is also quite well known and led to his creation of a fairly general theory of "dislocations."

Vito Volterra

At the outbreak of World War I, Volterra and others organized meetings and distributed propaganda urging Italy to enter the war on the side of the Allies. When Italy did so, Volterra enlisted in the armed forces, joining the air force at the age of 55. He established the Office of War Inventions in Italy and traveled frequently to England and France in order to help promote technical and scientific cooperation among the Allies.

After the war, Volterra resumed his position at the university in Rome. His most important work after 1918 was in the field of mathematical biology. Volterra investigated in great detail complex models for the interaction of species. We have seen that the Lotka-Volterra model predicts the existence of periodic fluctuations in the predator and prey species. Ecologists had previously observed such fluctuations, but had generally believed them to be explained only by external causes.

Volterra was drawn into studying this problem by a request from his son-in-law. Volterra's daughter Luisa had married her thesis advisor, a young marine biologist Umberto D'Ancona. D'Ancona had been examining data from the fish markets in Italian cities on the Adriatic between 1914 and 1923, a decade that included World War I. "He asked me," Volterra recalled, "if it were possible to give a mathematical explanation of the results he was getting in the percentages of the various species in these different periods." Volterra's subsequent research in mathematical models in biology resulted in more than 30 papers and books.

Volterra rebelled against the anti-Semitic Fascist government of Mussolini that held power in Italy from the 1920s until the early years of World War II. When Volterra refused to take an oath of allegiance to the government, he was stripped of his university position in 1931 and forced the next year to resign from all Italian scientific academies. He continued his mathematical research nonetheless and published continuously until shortly before his death on October 11, 1940. His published papers, numbering nearly 300, have been collected in five large volumes.

In her introduction to Volterra's biography, Judith R. Goodstein [2007] notes the situation 3 years after his death:

> *Volterra died in such obscurity that when the Nazis occupied Rome in 1943, German soldiers knocked on the door of his house . . . confidently expecting to arrest him and deport him to a concentration camp.*
>
> *Volterra's life exemplified the post-unification rise of Italian mathematics, its prominence in the first quarter of the 20th century, and its precipitous decline under Mussolini. This intellectual history in turn parallels the rise of Italian Jewry in the latter half of the 19th century and its travails during the Second World War. The meteoric rise and tragic fall of Volterra and his circle thus constitutes a lens through which we may examine in intimate detail the fortunes of Italian science in an epic scientific age.*

EXERCISES

II. Two Real-World Situations

1. Is the relation between a parasite and its host the same as that between predator and prey?

2. What assumptions underlie the conclusion that frequency of encounters between two species is proportional to the product of the two populations? Are these assumptions reasonable?

3. Show that the system $dG/dt = aG - bGL$, $dL/dt = mGF - nL$ reflects the three verbal assumptions made about the predator-prey situations.

4. Formulate another model that is consistent with the verbal assumptions about predator-prey situations.

5. Show that the system $dU/dt = aU - bUV$, $dV/dt = mV - nUV$ reflects the verbal assumptions made about the competitive hunters situation. Formulate a different model consistent with these assumptions.

III. Autonomous Systems

6. Show that $y = 0$, $x = e^{at}$ is a solution of both predator-prey and competitive hunters models. What do the orbits look like? In which direction are they traced out? Answer the same questions if it is specified that $x = 0$ is one of the functions in a solution.

7. Show that the functions F, G, F_x, F_y, G_x, G_y are continuous over the entire xy-plane in the cases where F and G come from

(a) Richardson arms race model

(b) Predator-prey model

(c) Competitive hunters model

8. Verify the details of the claims related to the example presented immediately before Theorem 2.

9. Show that in general, the Richardson arms race system has only a single one-point orbit.

10. Can the predator-prey or competitive hunters models fail to have two distinct one-point orbits? Why, or why not?

11. A critical point S of an autonomous system is called an *isolated critical point* if there is a circle of positive radius centered at S inside of which there are no other critical points.

(a) Show that the critical points of the predator-prey and competitive hunters models are isolated.

(b) Find an autonomous system with a nonisolated critical point.

12. An isolated critical point S of an autonomous system is *stable* if, given any positive number ε, there is a positive number δ such that (1) every orbit in the δ-neighborhood of S for some $t = t_1$ is defined for all $t > t_1$ and (2) if a trajectory satisfies (1), it remains in the ε-neighborhood of S for $t > t_1$.

Are the critical points of the ecological models that are presented in this chapter stable?

13. A stable critical point is called *asymptotically stable* if every orbit satisfying (1) and (2) of Exercise 12 also satisfies $\lim_{t\to\infty} (x(t), y(t)) = S$.

Are the critical points of the ecological models presented in this chapter asymptotically stable?

14. Which cases of the Richardson arms race model exhibit stable equilibrium? unstable equilibrium?

15. Check whether $x = y = e^{-t}$ is a solution of the system $x' = -x, y' = -y$. Can you find any other solutions?

IV. The Competitive Hunters Model

16. Can Euler's method be applied to analyze this model? What are the results?

17. Find a linearized Taylor series expansion for each of the following functions about the indicated point:

(a) $F(x, y) = y/x$ about (1, 2)

(b) $F(x, y) = \sin(xy)$ about (0, 0)

(c) $F(x, y) = \sqrt{y} + \log x$, about (4, 1)

18. If $G(x, y) = my - nxy$, show that the Taylor series approximation gives

$$G\left(\frac{m}{n} + h, \frac{a}{b} + k\right) \approx -\frac{an}{b}h$$

19. Use linearized Taylor series to study the nature of orbits of the competitive hunters model near the critical point (0, 0).

20. Consider the function $f(x) = x^m e^{-nx}$, where m and n are positive constants, and suppose that the domain of f is the set of all nonnegative numbers.

(a) Show that $f(x) \geq 0$ for all $x \geq 0$.

(b) Show that $f(x) = 0$ if and only if $x = 0$.

(c) Use l'Hôpital's rule to determine $\lim_{x\to\infty} f(x)$.

(d) By consideration of the first derivative, show that f has a maximum value when $x = m/n$. What is the maximum value?

(e) Find the points of inflection and regions of positive and negative concavity in the graph of f.

(f) Sketch a careful graph of f.

21. Choose numerical values for the parameters in the competitive hunters model, and use the Volterra mapping technique to locate at least a dozen points on an orbit of a solution.

22. The competitive hunters model is approximated, near the critical point $(m/n, a/b)$ by the simpler system $u'(t) = -(bm/n)v, v' = -(an/b)u$. Compute $u''(t)$ and $v''(t)$ and solve the resulting second-order differential equations to find exact solutions for u and v as functions of t. (This problem requires knowledge of differential equations beyond that demanded in the text.)

23. The equation $am^2(x - \frac{m}{n})^2 - b^2m(y - \frac{a}{b})^2 = K$ does not represent a hyperbola if $K = 0$.

(a) What does it represent?

(b) Are there initial levels of population (x_0, y_0) that would make $K = 0$?

(c) If $K = 0$, show that the orbit asymptotically may approach the critical point. Does this contradict the principle of competitive exclusion?

V. The Predator-Prey Model

24. Verify the details of the linearized Taylor series expansion for the predator-prey model.

25. Check whether $z = A\sin\sqrt{p}t + B\cos\sqrt{p}t$ satisfies $z'' = -pz$.

26. What can you say about the nature of the orbits for the predator-prey model if initial population levels make the constant $K = 0$ (Fig. 4.12)?

27. Carry out the indicated integration to verify that the average values of u and v are 0.

28. Use the Volterra mapping technique to graph some orbits of the predator-prey model if $a = 4$, $b = 2$, $m = 3$, and $n = 1$.

29. The orbits of the competitive hunters model are the graphs of solutions to the first-order differential equation

$$\frac{dy}{dx} = \frac{my - nxy}{az - bxy}$$

while orbits of the predator-prey model are graphs of solutions to the equation

$$\frac{dy}{dx} = \frac{mxy - ny}{ax - bxy}$$

By consideration of d^2y/dx^2 discuss whether the orbits have inflection points.

30. (*Kemeny and Snell*) The predator-prey and competitive hunters models are special cases of the more general model

$$dx/dt = xG(y)$$
$$dy/dt = yH(x)$$

where $G(0)$ and $H(0)$ are nonzero.

(a) Find the critical points of such a system.

(b) Are there any one-point orbits? Straight-line orbits?

(c) Prove that if $x_0 > 0$ and $y_0 > 0$, then $x(t) > 0$ and $y(t) > 0$ for all t, where $x_0 = x(0)$ and $y_0 = y(0)$.

(d) By consideration of dy/dx, find an equation whose graph is the orbit of such a system.

(e) Prove that if the solution is periodic, of period T, then

$$\int_0^T H(x)dt = \int_0^T G(y)dt = 0$$

(f) Let (p, q) be a critical point in the first quadrant other than the origin. Prove that the approximate orbits near this point are hyperbolic if $G'(q)H'(p) > 0$, and are elliptic if $G'(q)H'(p) < 0$.

(g) Show that the behavior of the solutions on the axes and near the origin depends only on $G(0)$ and $H(0)$.

(h) Show that these results are consistent with the information obtained about the predator-prey and competitive hunters models.

SUGGESTED PROJECTS

1. Re-examine the analysis of the arms race model of Chapter 2 in the light of the mathematical techniques presented in this chapter. Now that you know more about autonomous systems, you ought to be able to say more. Can you?

2. How does Bell modify the predator-prey model to study the immune response to infections? What are his conclusions? (See References.)

3. Generalize the competitive hunters model to reflect the assumption that one species experiences logistic population growth in the absence of the other. Formulate the model as a system of differential equations. Analyze the model using the techniques of this chapter. What are the conclusions? Are they consistent with observed behavior?

4. Analyze mathematically one or more of the three suggested variations of the predator-prey model. How do the conclusions differ from those of the simpler model?

5. Consider an ecological system with three interacting species that contains both predator-prey and competition as features. Formulate a system of differential equations to model this situation. Analyze the mathematical system and interpret the conclusions.

6. If the constants b and n are taken to be negative in the competitive hunters model, the resulting model represents what ecologists label *mutualism* or *symbiosis*. This is a relationship in which both species gain from their association with each other. It is a relationship favored by natural selection and is very common in nature. Find some instances of mutualism. Analyze the mathematical model. Interpret the results.

7. Modify the predator-prey model by assuming that in the absence of predators, the growth rate of the prey is constant instead of being proportional to the prey population—that is, $dx/dt = a - bxy$. Analyze this variation using the techniques developed in this chapter.

8. Investigate the field of "mathematical vampirology." Examine how the Lotka-Volterra model is the starting point for investigation such questions as "What is the optimal bloodsucking policy for the vampires?" and "How much of human labor should go for the production of stakes to kill vampires?" See the papers by Richard Hartl, Gustav Feichtinger, Alexander Mehlmann, Andreas Novak, and Dennis Snower listed below.

9. Examine Goodwin's [1967] *Class Struggle Model* of the relationship between wage-earning workers and profit-earning capitalists where cyclical behavior is captured. High employment generates wage inflation, which can increase the wage share of workers in output, lowering the profits of capitalists and reducing future investment and output. That reduction in output will in turn decrease labor demand and employment, consequently driving down the wage share of workers. But as wage share declines, then profits and hence investment increase, leading to greater employment. Then the bargaining power of workers improves and consequently their wages—and the rest of the cycle then repeats itself. Determine how the additional features Goodwin considers (e.g., growth in labor supply and improved productivity) affect the results of the Lotka-Volterra framework with which he began.

10. Analyze how Joshua Epstein [1997] expands on the Lotka-Volterra model to analyze political unrest and the spread of revolutionary ideas. Critically examine how he handles what he describes as (a) explosive upheavals, (b) revolutions that fizzle for lack of a receptive population, (c) revolutions that spread but that are reversed and crushed by an elite, (d) longer-term cycles of revolutionary action, (e) endemic levels of social discontent, and (f) traveling waves of revolution. What

features of revolutionary action does he omit? How might they be reflected in a revised model?

11. Analyze the Rosenzweig-MacArthur version of Harrison's model using the techniques we've developed in this chapter. Consider other choices for the response function $f(x)$ and analyze the resulting models.

12. Formulate and study a discrete dynamical version of one of the competition or predator-prey models. Do the discrete and the continuous models produce the same qualitative results? Are there values for the parameters that lead to chaos?

13. Analyze the behavior of the predator-prey or competitive hunters models if one or more species is subject to a strong Allee effect.

14. Carle Runge (1856–1927) and Martin Kutta (1867–1944) developed numerical methods for approximating solutions of systems of differential equations. One of the most popular is the *Runge-Kutta Method of Order 4*. It generates approximate points for the orbit of an autonomous system of differential equations

$$x'(t) = F(x, y),\ x(0) = x_0$$
$$y'(t) = G(x, y),\ y(0) = y_0$$

by the computations

$$h_1 = F(x_i, y_i)\Delta t \qquad k_1 = G(x_i, y_i)\Delta t$$

$$h_2 = F\left(x_i + \frac{h_1}{2}, y_i + \frac{k_1}{2}\right)\Delta t \qquad k_2 = G\left(x_i + \frac{h_1}{2}, y_i + \frac{k_1}{2}\right)\Delta t$$

$$h_3 = F\left(x_i + \frac{h_2}{2}, y_i + \frac{k_2}{2}\right)\Delta t \qquad k_3 = G\left(x_i + \frac{h_2}{2}, y_i + \frac{k_2}{2}\right)\Delta t$$

$$h_4 = F(x_i + h_3, y_i + k_3)\Delta t \qquad k_4 = G(x_i + h_3, y_i + k_3)\Delta t$$

$$h = \frac{h_1 + 2h_2 + 2h_3 + h_4}{6} \qquad k = \frac{k_1 + 2k_2 + 2k_3 + k_4}{6}$$

Then let $\begin{aligned} x_{i+1} &= x_i + h \\ y_{i+1} &= y_i + k \end{aligned}$.

Use the Runge-Kutta technique to plot orbits for the competition and predator-prey models. Investigate the accuracy of the Runge-Kutta method as compared to the Euler method. What is the geometric basis for the Runge-Kutta approach? Most differential equations texts will provide an introduction.

You can find a listing of references and suggestions for additional reading on the book's website, www.wiley.com/college/olinick

Tumor Growth Models

CHAPTER 5

> What we call growth of even a simple organism is a tremendously complex phenomenon from the biochemical, physiological, cytological and morphological viewpoints. There are, however, certain aspects that are amenable to quantitative analysis, and such an approach appears to lead to some insight into the connections between metabolism and growth, and to some answer to the seemingly trivial, but in fact hardly explored question, "Why does an organism grow at all, and why, after a certain time, does its growth come to a stop?"
>
> —Ludwig von Bertalanffy

I. Introduction

Cancer is a leading cause of cause of death worldwide, accounting for nearly 8 million deaths per year. Experts predict that deaths globally will continue rising, with an estimated 9 million people dying from cancer in 2015 and 11.4 million succumbing in 2030. Thus, cancer poses major public health questions.

Cancer is a generic term for a large class of diseases that can affect any part of the body; *malignant tumors* and *neoplasms* are often used as synonyms. A characteristic feature of cancer is the rapid creation of abnormal cells growing beyond their usual boundaries and often invading adjoining parts of the body, spreading to other organs. Many cancers can be detected early and powerful treatments exist. Yet our understanding of cancer is far from complete, and a diagnosis of cancer often results in great anxiety and fear for the patient. The writer Susan Sontag has noted how frequently words such as "horror" and "dread" are associated with cancer. In her book *Illness as Metaphor*, she argues that cancer is often seen "as no mere disease but a demonic enemy":

> *Two diseases have been spectacularly, and similarly, encumbered by the trappings of metaphor: tuberculosis and cancer. The fantasies inspired by TB in the last century, by cancer now, are responses to a disease thought to be intractable and capricious—that is, a disease not understood—in an era in which medicine's central premise is that all diseases can be cured. Such a disease is, by definition, mysterious. For as long as its cause was not understood and the ministrations of doctors remained so ineffective, TB was thought to be an insidious, implacable*

theft of a life. Now it is cancer's turn to be the disease that doesn't knock before it enters, cancer that fills the role of an illness experienced as a ruthless, secret invasion—a role it will keep until, one day, its etiology becomes as clear and its treatment as effective as those of TB have become.

Understanding the dynamics of cancerous tumor growth may help develop better prognoses for patients and more effective treatment plans. In this chapter, we will investigate several mathematical models of tumor growth that have proved effective in promoting knowledge about cancer. We also introduce the important *method of least squares* for fitting a model to observed data by determining appropriate values for the model's parameters.

An ideal mathematical model for a real-world situation should satisfy several criteria:

- The model should have a basis in reality.
- The model should have a minimum number of parameters.
- Variables represented in the model should be measurable so that it is possible to collect experimental data.
- The model's predictions should be reasonably accurate and give a good fit to experimental data.
- The model should improve our understanding of the real-world situation.

To these criteria, Vinay Vaidya and Frank Alexandro Jr. add several more desirable features for an ideal *model of tumor growth*:

- The model should have a physiological basis.
- The model should improve general understanding at microscopic as well as macroscopic level of tumor growth.
- The model should have breadth, in the sense that it should be applicable to different patients or animals with the same type of tumor.

Since more than 100 different diseases fall under the label of cancer, it would not be reasonable to expect that a single mathematical model would represent well all the diverse tumors that can beset some many different parts of the body. We will begin with a generalized "two-parameter" model, move on to examine a special limiting case of this generic model, and then focus on some models of a particular type of cancerous tumor.

II. A General Tumor Growth Model

The size of a dynamically changing entity (be it a single cell, tumor, urban population or economy) depends on a rate of increase and a rate of loss. In biology, terms such as *proliferation* and *synthesis* indicate growth. Words like *death* or *degradation* describe loss. Scientists have often employed *allometric* or *power law* models in which rates of change of

the size of a biological organism are proportional to powers of the size. In the mid-20th century, Ludwig von Bertalanffy proposed a general form for such models,

$$\frac{dy}{dt} = ay^{\alpha} + by^{\beta} \tag{1}$$

where y is a measure of the size of the organism and a, b, α, and β are constants. Eq. (1) is called a *generalized Bertalanffy model* with parameters α and β.

For tumors, size can be measured by volume, biomass, or number of cells. For convenience, we will use volume (V) as the measure of size. Taking the first term on the right-hand side of Eq. (1) as representing increase and the second term for loss, we can rewrite that equation as

$$\frac{dV}{dt} = aV^{\alpha} - bV^{\beta} \tag{2}$$

where a and b are positive.

Note that the special case $\alpha = 1$, $\beta = 2$ gives the logistic equation

$$\frac{dV}{dt} = aV - bV^2, V(0) = V_0 \tag{3}$$

we studied in Chapter 3 and that has the solution

$$V(t) = \frac{aV_0}{bV_0 + e^{-at}(a - bV_0)} = \frac{a}{b}\left[1 - \left(1 - \frac{a}{bV_0}\right)e^{-at}\right]^{-1} \tag{4}$$

Von Bertalanffy himself was quite interested in the relationship between body size (as measured by weight) and metabolic rate. While weight is directly proportional to volume, metabolic rate seemed to be proportional to surface area. Since volume is measured in cubic units of length and area in square units, surface area is proportional to the two-thirds power of volume.

"Animal growth," von Bertalanffy observed, "can be considered as a result of a counteraction of synthesis and destruction, of the anabolism and catabolism of the building materials of the body. There will be growth so long as building up prevails over breaking down." Experimental evidence indicated that the building up process was largely tied to metabolism while catabolism, the loss of building material, is directly proportional to weight and hence directly proportional to volume. These considerations led von Bertalanffy to investigate the particular version of Eq. (1) in which case $\alpha = 2/3$ and $\beta = 1$:

$$\frac{dV}{dt} = aV^{2/3} - bV, V(0) = V_0 \tag{5}$$

Fig. 5.1 shows the direction field for the Bertalanffy equation. Note that it shares some of the qualitative features as the direction field for the logistic equation.

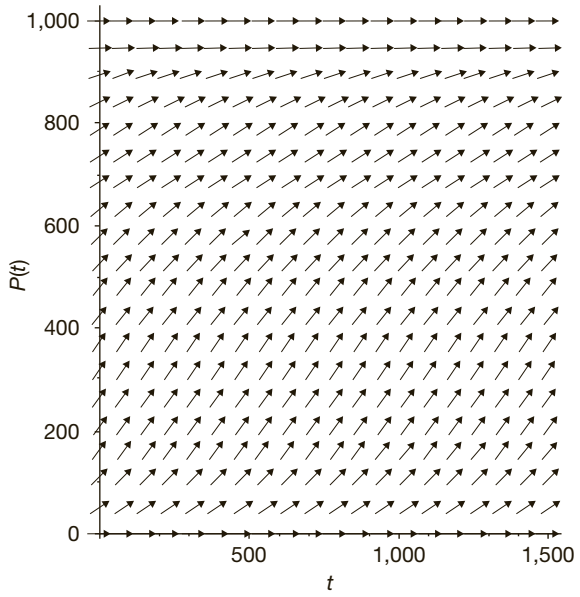


FIGURE 5.1 Direction field for Bertalanffy equation.

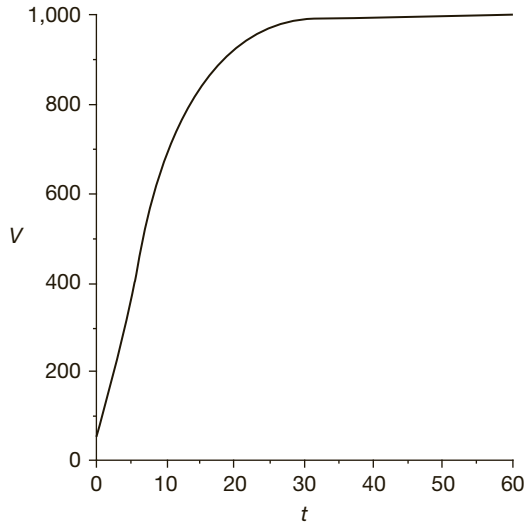


FIGURE 5.2 The graph of the solution of the von Bertalanffy curve for $a = 5$, $b = ½$, and $V_0 = 50$. Note that the $\lim_{t\to\infty} V(t) = 1000$.

We can solve Eq. (5) by separation of variables and integration to obtain

$$V(t) = \left[\frac{a}{b} - e^{-\frac{bt}{3}}\left(\frac{a}{b} - V_0^{\frac{1}{3}}\right)\right]^3 \tag{6}$$

In Fig. 5.2, we display the graph of Eq. (6) for a particular choice of a, b, and V_0.

The logistic and von Bertalanffy equations have both been used successfully to model the growth of some tumors in mice. See V. G. Vaidya and F. J. Alexandro Jr., "Evaluation of Some Mathematical Models for Tumor Growth," *International Journal of Bio-medical Computing* **13**(1982): 19−35.

III. The Gompertz Model

In 1825, the English mathematician Benjamin Gompertz proposed a new model for certain growth processes that scientists in many disciplines have found to be a valuable tool. We will examine two different lines of thought that lead to this model.

A. Introducing the Model

First, consider an expression of the form

$$AV^{\alpha} + BV^{\alpha}\left[\frac{V^{x}-1}{x}\right]$$

and replace B by bx and A by $a + B/x$ to rewrite this expression as

$$\left(a + \frac{B}{x}\right)V^{\alpha} + bxV^{\alpha}\left[\frac{V^{x}-1}{x}\right] = aV^{\alpha} + \frac{B}{x}V^{\alpha} - bV^{\alpha} + bV^{\alpha+x} \tag{7}$$

which, since $B/x = b$, simplifies to $aV^{\alpha} + bV^{\beta}$, where $\beta = \alpha + x$.

Hence, the differential equation

$$\frac{dV}{dt} = AV^{\alpha} + BV^{\alpha}\left[\frac{V^{x}-1}{x}\right] \tag{8}$$

is equivalent in form to the generalized Bertalanffy equation $\dfrac{dV}{dt} = aV^{\alpha} + bV^{\beta}$.

Now we consider the limiting case of Eq. (8) when $x \to 0$. By l'Hôpital's Rule,

$$\lim_{x\to 0}\left[\frac{V^{x}-1}{x}\right] = \lim_{x\to 0}\frac{V^{x}\ln V - 0}{1} = \ln V \tag{9}$$

and Eq. (8) becomes

$$\frac{dV}{dt} = AV^{\alpha} + BV^{\alpha}\ln V \tag{10}$$

In the particular case in which $\alpha = 1$, we have

$$\frac{dV}{dt} = AV + BV\ln V \tag{11}$$

which is called the *Gompertz equation*. Eq. (10) is called the *generalized Gompertz equation*. Thus, the generalized Gompertz equation may be seen as a limiting case of the generalized Bertalanffy equation.

Setting $A = a$ and $B = -b$ where a and b are positive, we may write the Gompertz equation as

$$\frac{dV}{dt} = aV - bV\ln V = V(a - b\ln V) \tag{12}$$

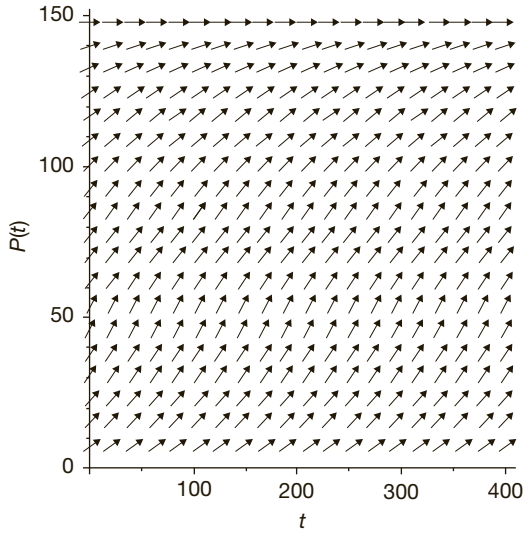


FIGURE 5.3 Direction field for Gompertz curve.

Observe that Eq. (12) implies that the time derivative dV/dt is zero when $V = 0$ and when $V = e^{a/b}$. Note also that this derivative is positive for $0 < V < e^{a/b}$ and is negative when $V > e^{a/b}$.

Fig. 5.3 displays the direction field for the Gompertz equation.

B. Solving the Gompertz Equation

The Gompertz differential equation is easy to solve by separating the variables and integrating:

$$\int \frac{1}{V(a - b \ln V)} dV = \int 1 \, dt \tag{13}$$

Make the change of variable $u = a - b \ln V$ so $du = -\dfrac{b}{V} dV$, and our integral problem becomes

$$\int \frac{-1}{bu} du = \int 1 \, dt \text{ or } \int \frac{1}{u} du = \int -b \, dt \tag{14}$$

and hence,

$$\ln u = -bt + C \tag{15}$$

which gives us

$$u = Ce^{-bt} \tag{16}$$

or, in terms of our original variables,

$$a - b \ln V = Ce^{-bt} \tag{17}$$

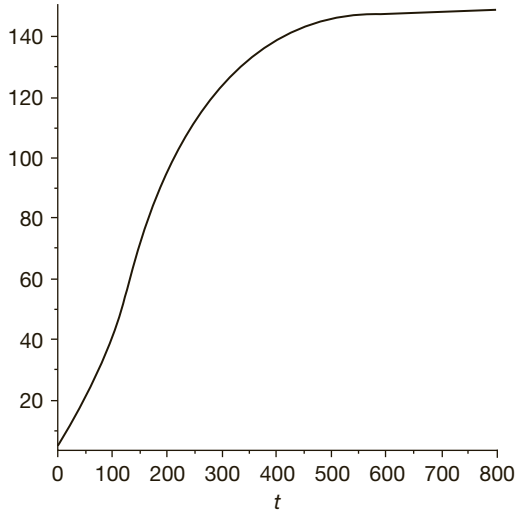


FIGURE 5.4 A graph of the Gompertz curve.

Since $V(0) = V_0$, we have

$$C = a - b \ln V_0$$

Solving Eq. (17) for V, we obtain

$$\begin{aligned} \ln V &= \frac{a - Ce^{-bt}}{b} \\ V &= e^{\frac{a - Ce^{-bt}}{b}} \\ V &= e^{\frac{a - (a - b \ln V_0)e^{-bt}}{b}} = e^{\frac{a}{b} - \left(\frac{a}{b} - \ln V_0\right)e^{-bt}} \end{aligned} \tag{18}$$

Fig. 5.4 displays a Gompertz curve, the graph of the function in Eq. (18).

C. An Alternative Derivation

It's useful to examine a different derivation of the Gompertz equation, one that is motivated by a consideration that growth occurs within environments that have limited natural resources.

The logistic model we studied in Chapter 3 provides one way to represent the fact that biological environments have limited resources that make sustained exponential growth impossible. A chief characteristic of the logistic model is the idea of *carrying capacity*, a maximum sustainable population size.

The Gompertz model addresses limited resources from a different perspective. This model takes into account that even in a resource-rich environment, the amount of resources

available to an individual cell in a tumor may depend on the cell's *location* within the tumor. Cells on the outer surface of the tumor have better access to the oxygen and nutrients necessary for replication to continue, so they are more fit than the cells at the interior of the tumor.

If we assume that a tumor is roughly in the shape of a sphere, then the surface area-to-volume ratio is $\frac{4\pi r^2}{\frac{4}{3}\pi r^3} = \frac{3}{r}$, which decreases as a tumor grows. Thus, one would expect the growth rate of a tumor to decrease correspondingly as the tumor became larger.

Suppose that at every instant the size of the tumor is growing at a constant percentage rate, but that rate is also decreasing in a similar manner. Then we can describe the dynamics of tumor growth as a simple system of differential equations,

$$\begin{cases} \dfrac{dV}{dt} = r(t)V \\ \dfrac{dr}{dt} = -kr \end{cases} \tag{19}$$

where k is a positive constant. From the second equation of (19), we have $r(t) = r_0 e^{-kt}$. Then the first equation of (19) becomes

$$\frac{dV}{dt} = r_0 e^{-kt} V \tag{20}$$

which we can solve by separating the variables and integrating:

$$\int \frac{1}{V} dV = \int r_0 e^{-kt} dt \tag{21}$$

$$\ln V = -\frac{r_0}{k} e^{-kt} + C \text{ for some constant } C \tag{22}$$

Solving for V as a function of t, we obtain

$$V = Ce^{-\frac{r_0}{k}e^{-kt}} \text{ for some constant } C \tag{23}$$

If $V(0) = V_0$, then we have $V_0 = Ce^{-\frac{r_0}{k}}$ so $C = V_0 e^{\frac{r_0}{k}}$, and the solution of this form of the Gompertz equation becomes

$$V = V_0 e^{\frac{r_0}{k}} e^{-\frac{r_0}{k}e^{-kt}} = e^{\ln V_0} e^{\frac{r_0}{k}} e^{-\frac{r_0}{k}e^{-kt}} = e^{\frac{r_0}{k} - \frac{r_0}{k}e^{-kt} + \ln V_0} \tag{24}$$

Although the forms of the two versions of the Gompertz model and their solutions appear somewhat different, it is not difficult to verify that they are equivalent. One way to see this equivalence is to let $k = b$ and $r_0 = a - b \ln V_0$ in Eq. (24). These substitutions convert Eq. (24) to

$$V = e^{\frac{a-b\ln V_0}{b} - \frac{a-b\ln V_0}{b}e^{-bt} + \ln V_0} = e^{\frac{a}{b} - \ln V_0 + -\ln V_0 - \frac{a-b\ln V_0}{b}e^{-bt}} = e^{\frac{a}{b} - \frac{a-b\ln V_0}{b}e^{-bt}}$$

which is identical to Eq. (18).

D. Estimating the Parameters

Suppose we plot data on the growth of some organism over time and observe that the graph *appears* to have the shape of a Gompertz curve. How can we estimate the values of the parameters of a Gompertz curve? How closely do these parameters predict the observed data?

We begin with the solution of the Gompertz equation in the form $V(t) = V_0 e^{\frac{r_0}{k}} e^{-\frac{r_0}{k}e^{-kt}}$. Assuming that our first measurement V_0 occurs at time $t = 0$, we want to make the best estimates for r_0 and k. We'll begin by taking the logarithms of each side of this equation for $V(t)$:

$$\ln V(t) = \ln V_0 + \ln e^{\frac{r_0}{k}} + \ln e^{-\frac{r_0}{k}e^{-kt}} \tag{25}$$

so

$$\ln V(t) = \ln V_0 + \frac{r_0}{k} - \frac{r_0}{k}e^{-kt}$$

and hence (replacing t by $t - 1$)

$$\ln V(t-1) = \ln V_0 + \frac{r_0}{k} - \frac{r_0}{k}e^{-k(t-1)}.$$

If we take the difference of these last two expressions, we have

$$\ln V(t) - \ln V(t-1) = -\frac{r_0}{k}e^{-kt} + \frac{r_0}{k}e^{-k(t-1)} = e^{-kt}\left(\frac{r_0}{k}e^{k} - \frac{r_0}{k}\right) \tag{26}$$

which has the form

$$W(t) = Ae^{-kt}$$

where $W(t) = \ln V(t) - \ln V(t-1)$ and $A = \left(\frac{r_0}{k}e^{k} - \frac{r_0}{k}\right) = \frac{r_0}{k}\left(e^{k} - 1\right)$.
Now if examine the logarithm of $W(t)$, we have

$$Y(t) = \ln W(t) = \ln A - kt \tag{27}$$

so the graph of Y as a function of t is a straight line.

If the growth of V is indeed determined by a Gompertz curve, then we can find A and k if we can find the coordinates of any two points on this straight line. This is possible if we know the values of V for any three consecutive equally spaced times: t, $t - 1$, and $t - 2$. Once we know A, we can determine r_0 as $r_0 = \frac{Ak}{e^{k} - 1}$. We followed a similar approach in fitting a logistic curve to U.S. census data in Chapter 3. But real-world measurements are seldom, if ever, exact. Errors in experimental measurement are the rule, not the exception. Thus, a different triple of consecutive values for V might well yield a different set of values for k and r_0. Is there a different approach to obtain a *best* set of values for the parameters k

and r_0? An oft-used technique is to choose the parameter values that minimize the total difference between measured values and predicted values.

If we have a set of measured values $Y_1, Y_2, \ldots, Y_N$ corresponding to times $1, 2, \ldots, N$ that graphically appear to lie along a straight line $y = a - kx$, and we wish to determine the "best" choices for a and k, then the difference between the observed value and the predicted value at time i would be $a - ki - Y_i$. There are several ways we can measure the cumulative fit between observations and predictions. These include

$$\sum_{i=1}^{N} (a - ki - Y_i) \text{ or } \sum_{i=1}^{N} \left| a - ki - Y_i \right| \text{ or } \sum_{i=1}^{N} (a - ki - Y_i)^2$$

The first sum adds up all the differences and seems the most natural measure to use, but this sum might turn out to be very small—not because all the individual differences are small, but because some very large positive differences are balanced out by equally large negative differences. In such a case, all our individual predictions might be considerably far from the observed values. The second sum, which takes the absolute values of the "error" terms, avoids this problem. The second sum can be small only when all the individual terms are small. It's a much better sum to use, but suffers from the fact that it is analytically rather difficult to minimize a sum of absolute values. The third sum is easier to deal with using the tools of calculus and also has the property that it is small only when all the individual terms are small. Note also that summing the squares of the differences between predicted and observed values gives more weight to any predictions that are significantly different from the measured values.

E. Method of Least Squares

Minimizing the sum of squares of differences is called the *method of least squares*. In this method, we want to choose values for a and k that produce the smallest value of the *least squares function*

$$f(a, k) = \sum_{i=1}^{N} (a - ki - Y_i)^2 \tag{28}$$

There are several ways to determine these best values of a and k. We'll illustrate a calculus-based approach here, using a theorem that states that the minimum value of f will occur at choices for a and k that make both partial derivatives of f equal zero. See Appendix IV for more details.

In our case,

$$\frac{\partial f}{\partial a} = \sum_{i=1}^{N} 2(a - ki - Y_i) \text{ and } \frac{\partial f}{\partial k} = \sum_{i=1}^{N} (a - ki - Y_i)(-i) \tag{29}$$

The condition that both partial derivatives be 0 at a minimum gives the equations

$$\sum_{i=1}^{N} 2(a - ki - Y_i) = 0 \text{ and } \sum_{i=1}^{N} 2(a - ki - Y_i)(-i) = 0 \tag{30}$$

Table 5.1 The weight V_i in kilograms of a female chicken at the end t_i of week i. The data are taken with permission from Dragan Jukic, Gordana Kralik, and Rudolf Scitovski, "Least-Squares Fitting Gompertz Curve," *Journal of Computational and Applied Mathematics* **169** (2004): 359–375.

t_i	1	2	3	4	5	6	7	8	9	10	11	12	13
V_i	.147	.357	.641	.98	1.358	1.758	2.159	2.549	2.915	3.251	3.510	3.740	3.925

which simplify to

$$\sum_{i=1}^{N} a - \sum_{i=1}^{N} ki = \sum_{i=1}^{N} Y_i \text{ and } \sum_{i=1}^{N} -ai + \sum_{i=1}^{N} ki^2 = -\sum_{i=1}^{N} Y_i i \tag{31}$$

or

$$Na - \left(\sum_{i=1}^{N} i\right)k = \sum_{i=1}^{N} Y_i \text{ and } \left(-\sum_{i=1}^{N} i\right)a + \left(\sum_{i=1}^{N} i^2\right)k = -\sum_{i=1}^{N} Y_i i \tag{32}$$

Noting that $\sum_{i=1}^{N} i - \frac{N(N+1)}{2}$ and $\sum_{i=1}^{N} i^2 = \frac{N(N+1)(2N+1)}{6}$, the equations become

$$Na - \frac{N(N+1)}{2}k = \sum_{i=1}^{N} Y_i \text{ and } \frac{-N(N+1)}{2}a + \frac{N(N+1)(2N+1)}{6}k = -\sum_{i=1}^{N} Y_i i \tag{33}$$

But this is just a system of two linear equations in two unknowns (a and k), which is easily solved.

We illustrate this process with data collected by three Croatian scientists. Dragan Jukic, Gordana Kralik, and Rudolf Scitovski recorded the weight V in kilograms of a female chicken over a period of 13 weeks. Table 5.1 displays their results.

We'll let $t = 0$ correspond to the end of the week 1. Then $t = 1$ represents the time at the end of week 2, $t = 2$ the time at the end of week 3, and so on. These will give us values $V_0 = V(0) = .147$, $V(1) = .357$, . . . , $V(12) = 3.925$.

From our earlier discussion, the Gompertz model may provide a good fit for this data if the graph of $\ln W(t) = \ln(\ln(V(t)) - \ln(V(t-1)))$ as a function of time appears to fall along a straight line. Fig. 5.5 shows this graph; it does appear that the observed data produce points that seem to lie along a line

In this case, $f(a, k) = 39.88968422a - 330.9345822k + 650k^2 - 156ak + 12a^2 + 42.17943682$. The system of linear equations we need to solve is

$$1300k - 156a = 330.9345821$$
$$156k - 24a = 39.88968422$$

The solution is $k = .2505301912$ and $a = -0.03362393333$. Since the original measured data on chicken weight is accurate only to a limited number of decimal places, we will use

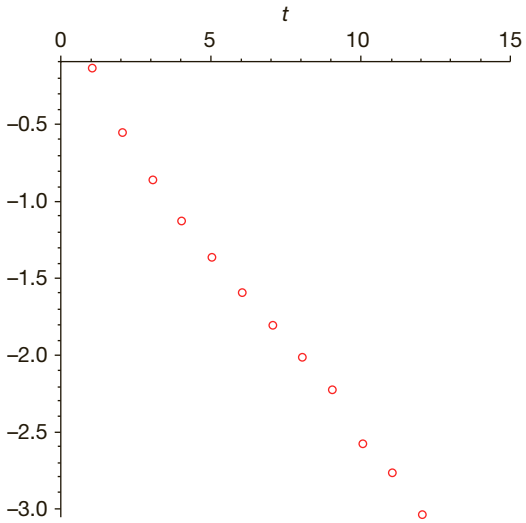


FIGURE 5.5 The graph of $\ln(W(t))$ for the chicken data of Table 5.1, where $W(t) = \ln(V(t)) - \ln(V(t-1))$.

Table 5.2 The observed and predicted values for the weights of the chickens obtained by the method of least squares. We have rounded off the predicted values to 3 decimal places, the accuracy of the observed values.

Week	Observed Value	Predicted Value	Predicted − Observed	Error Squared
1	0.147	0.147	0	0
2	0.357	0.313	0.044	0.002
3	0.641	0.564	0.077	0.006
4	0.98	0.892	0.088	0.008
5	1.358	1.274	0.084	0.007
6	1.758	1.683	0.075	0.006
7	2.159	2.090	0.069	0.005
8	2.549	2.474	0.075	0.006
9	2.915	2.821	0.094	0.009
10	3.251	3.125	0.126	0.016
11	3.51	3.384	0.126	0.016
12	3.74	3.600	0.140	0.020
13	3.925	3.778	0.147	0.021

rounded-off values $k = .25$ and $a = -0.03$. For these particular values of a and k, we have $f(a, k) = 0.05490074579$.

Recalling that $a = \ln A$, we have $A = e^a = e^{-0.03} = 0.97$. Thus, our best estimate for r_0 is $r_0 = \dfrac{Ak}{1 - e^k} = 0.854$. To complete our fitting of the data to a Gompertz function, note that $V_0 = .147$.

In our chicken example, our estimated Gompertz function would be $g(t) = V_0 e^{\frac{r_0}{k}} e^{-\frac{r_0}{k}e^{-kt}} = 0.147\mathrm{e}^{\frac{0.854}{.25}} e^{\frac{-0.854}{.25}e^{-0.25t}} = 4.47e^{-3.41e^{-.251\mathrm{t}}}$. Table (5.2) displays the observed and

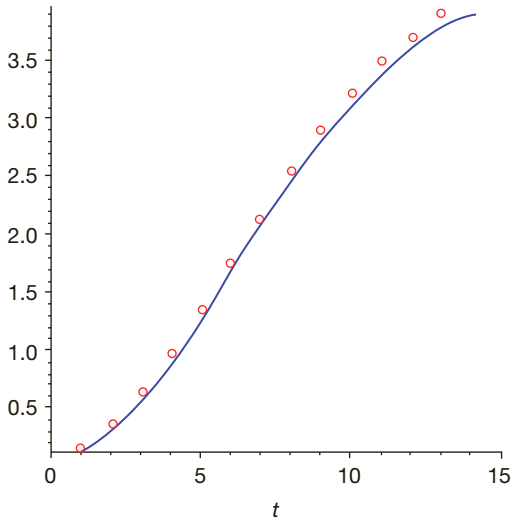


FIGURE 5.6 Comparison of the Gompertz model (solid curve) with the observed data (circled points) on chicken weights found in Table 5.1.

predicted values for the weights of the chickens and the differences between these numbers. Fig. 5.6 presents the observations and model predictions in visual form. It certainly appears from the graph in Fig. (5.6) that the Gompertz function we have found does an excellent job of matching the observed data.

Is there a quantitative way of measuring how well the model first the real-world data? In Chapter 3, we looked at the percentage errors the model makes and declared the model a good one if the largest percentage error was small. A more commonly used indicator of "goodness of fit" is the sum of the squared errors, where we add up the squares of the differences between the observed values and the model's predicted values. For our chicken example, the observed values are the V_t's and the predicted values are the $g(t)$'s. Our measure of goodness of fit is $\sum_{t=1}^{13} [V_t - g(t)]^2$. In this example, that sum is 0.121. We can then compare two different choices of parameter values (V_0, r_0, k) for the Gompertz models by seeing which one gives a smaller sum of squared errors.

We may also use the method of least squares to determine which of several different models for the same situation gives a better fit to the observed data. For example, we could ask how the Gompertz model for the chicken example compares with the logistic one.

Let's fit a logistic curve $V = \dfrac{k}{1 + e^{d-at}}$ to the data for the chicken weights, using the same limiting value for the weight, 4.479; $V = \dfrac{4.479}{1 + e^{d-at}}$ for some parameter values d and a that we need to estimate.

Rearrange the terms $e^{d-at} = \dfrac{4.479}{V} - 1$.

Then take the logarithm of both sides: $d - at = \ln\left(\dfrac{4.479}{V} - 1\right)$.

Introduce the change of variables: $Z = \ln\left(\dfrac{4.479}{V} - 1\right)$. The previous equation becomes the linear relationship

$$Z = d - at.$$

We can proceed to use the method of least squares to estimate d and a. The analysis yields $d = 3.155450907$ and $a = 0.4124054532$. If we use these parameter values, then the graphs of the resulting logistic equation and the original data can be seen in Fig. 5.7. Visually, it appears that our "best" Gompertz model does a slightly better job of matching the observed data than our "best" logistic model. In fact, the sum of squared differences between predicted and observed values for this logistic model is 0.269, which is larger than the value of 0.121 we found for the Gompertz model. Therefore, the Gompertz model is a better description of the growth of a chicken's weight than the logistic model, at least for this particular data set.

Fig. 5.8 shows an example a tumor growth data for which the Gompertz approach clearly mirrors reality much better than the logistic one. Scientists have found that in many

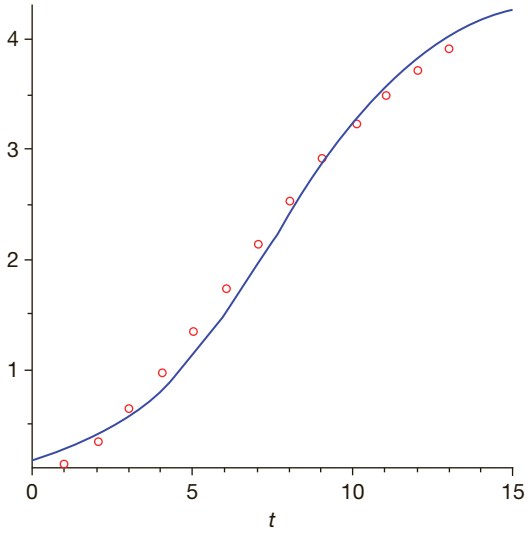


FIGURE 5.7 Comparison of the logistic model (solid curve) with the observed data (circled points) on chicken weight found in Table 5.2.

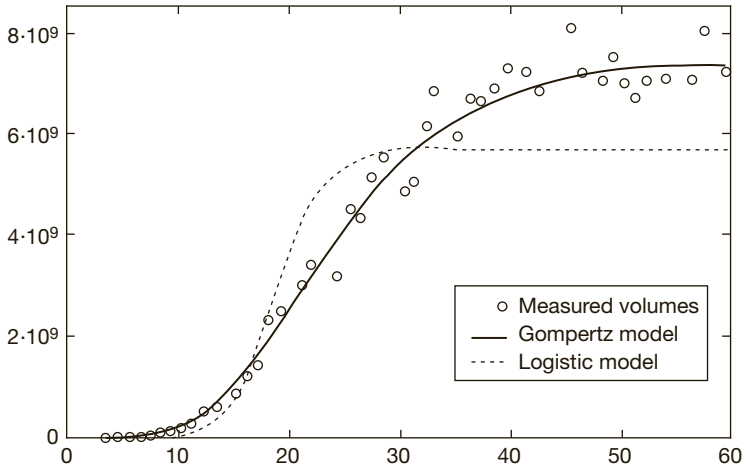


FIGURE 5.8 Best-fit curves by Gompertz and logistic models where the horizontal axis represents time in days and the vertical axis represents tumor volume. From Miljenko Marušić, "Mathematical Models of Tumor Growth," *Mathematical Communications* (Department of Mathematics), University of Osijek 1 (1996): 175–192.

tumor growth situations, the Gompertz model provides a more accurate description of the real-world observed data than the logistic model, but in other cases of tumor dynamics, the logistic fits the observations better than the Gompertz model. No single model type uniformly provides the best fit for every type of tumor.

In the next section, we will examine some models of tumor growth in the human colon that take into account particular features of this type of cancer. Before we do, it's useful to take a brief look at what's similar in the logistic, von Bertalanffy, and Gompertz models. These are all differential equations of the form $dx/dt = g(x)$, where g is a continuous function of x. One way to see qualitative similarities among these models is to examine the graph of g as a function of x in each case. Fig. 5.9 shows typical examples of these graphs. Note that all three curves are concave down, increase to a unique positive maximum, and then decrease.

The graphs of solutions to the Bertalanffy, logistic, and Gompertz models have a shape that resembles an elongated letter S. As the input variable increases, the output at first increases fairly slowly, then builds up more rapidly as the graph has a midsection that looks like a straight line before transitioning to an almost horizontally flat curve. The graphs of other functions with a similar shape are called *S-curves* or *sigmoid curves*.

The qualitative nature of an S-curve has been observed in the plotting of data in many fields, including demographics, biology, and economics. The growth of various components of an economy, the diffusion of new technologies and ideas, and the demand for new products often exhibit the three characteristics of an S-curve: emergence, inflexion, and saturation. The papers in the References by Gloria Jarne, Julio Sánchez-Chóliz, and Francisco Fátas-Villafranca examine the mathematical properties and applications of S-curves.

IV. Modeling Colorectal Cancer

While the Gompertz model often provides a good predictive model for how tumors grow in general, there are better models for some particular types of cancer that incorporate specific features of that growth. We present in this section several such models that shed light on colorectal cancer.

A. Colon Cancer

Colorectal cancer, also called *colon cancer* or *large bowel cancer*, includes cancerous growths in the colon, rectum, and appendix. It is the third most common form of cancer and is the second leading cause of cancer-related deaths in the Western world. Colorectal cancer causes 655,000 deaths worldwide per year. Many colorectal cancers are thought to arise from adenomatous polyps in the colon. These mushroom-like growths are usually benign, but some may develop into cancer over time.

Author David Guterson captures some of the individual human drama and the course of the disease of colon cancer in his novel *East of the Mountains*:

> *Dr. Ben Givens shrugged off his past to devote himself to the rain's steady cadence, but no dreams, no deliverance, came to him. Instead he . . . lay tormented by the unassailable fact that he was dying—dying of colon cancer. . . .*
>
> *Now he'd been told—it was the dark logic of the world—that he had months to live, no more. Like all physicians, he knew the truth of such a verdict; he knew full well the force of cancer and how inexorably it operated. He grasped that nothing could stop his death, no matter*

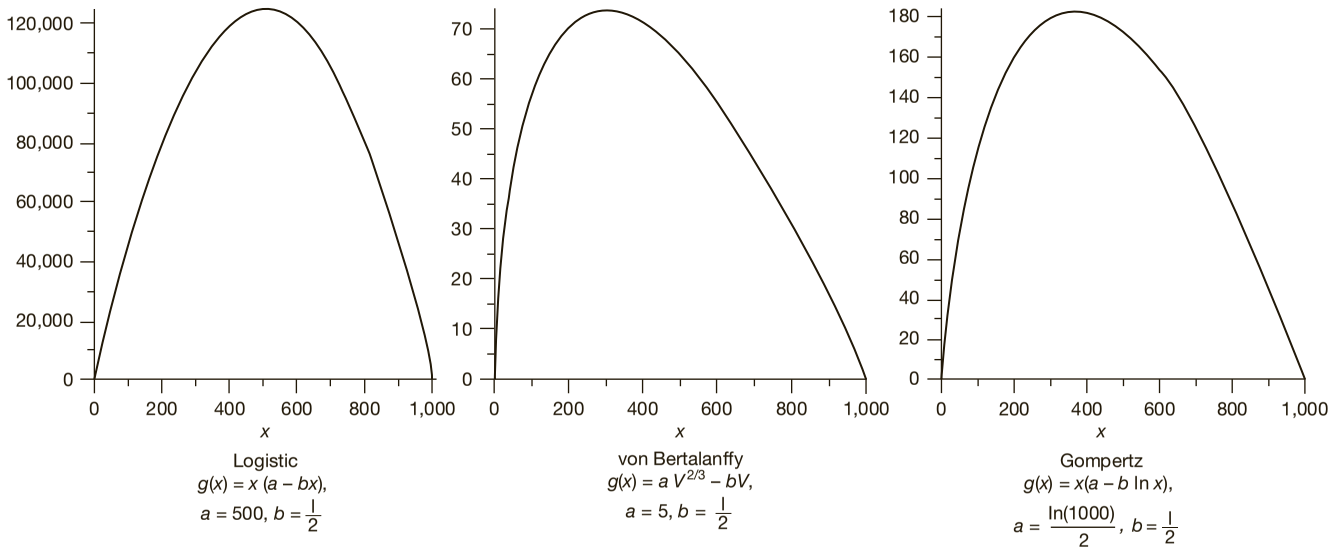
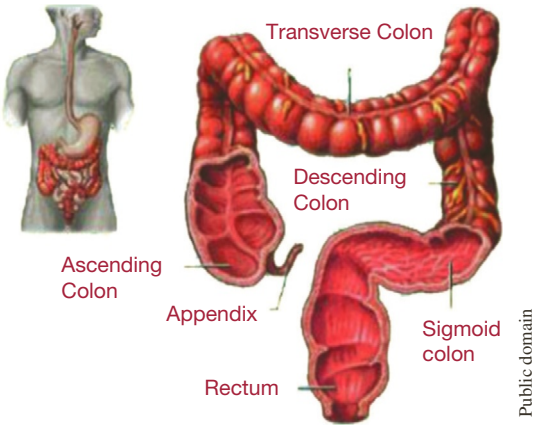


FIGURE 5.9 Graphs of $g(x)$ versus x for the logistic, von Bertalanffy, and Gompertz models $dx/dt = g(x)$.



Public domain

FIGURE 5.10 The location and major sections of the human colon. *Source:* www.moiracolonicclinic.com/images/colon.jpg.

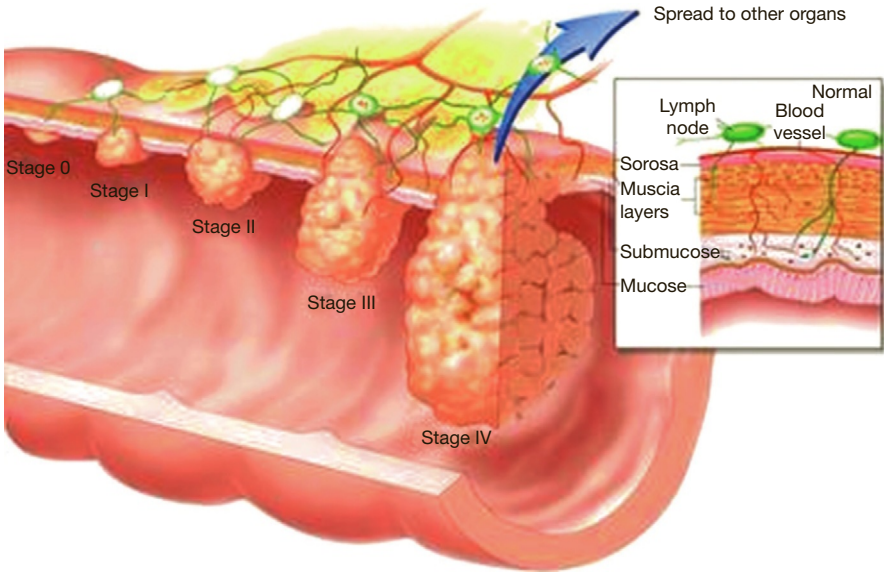
how hopeful he allowed himself to feel, no matter how deluded. Ben saw how his last months would be, the suffering that was inevitable, the meaningless trajectory his life would take into a meaningless grave. . . .

His cancer had metastasized, traveling from the mucosa of his colon to the lymph nodes closer to his tumor, and from there to sites in his liver. . . . He knew exactly what to expect and could not turn away from meeting it. After the bedsores and bone fractures, the bacterial infection from the catheter, the fluid accumulating between his viscera that would have to be expunged through a drainage tube; after the copious vomiting, the dehydration and lassitude, the cracked lips, dry mouth, tube feedings, and short breath, the dysphagia, pneumonia, and feverishness, the baldness and endgame sensation of strangling; after he had shrunk to eighty-five pounds and was gasping his last in a nursing-home bed. . . .

The colon is a common site for cancer. The reasons for a relatively high rate of colon cancer are not completely known; they may involve diet and/or the large quantities of new cells are produced there every day. A high replication rate results in a greater incidence of mutations, some of which may cause the impairment of control mechanisms and thus result in tumor growth. The physical organization of the colon into 10^7 *crypts* (invaginations in the lining of the colon), however, helps to prevent uncontrolled growth.

Each crypt contains stem cells at its base. This model assumes that a stem cell can divide to create either two new stem cells (a process called *regeneration* or *renewal*) or two differentiated cells. [Some biologists conjecture that a single stem cell may also evolve into one stem cell and one differentiated cell.] The differentiated cells migrate up the crypt, becoming progressively more specific as they reach the top, at which point they undergo *apoptosis* (programmed cell death). Each cell's journey from the bottom to the top of the crypt takes several days. Since each cell stays within its own crypt and has a relatively short life cycle, any mutation is not likely to become part of a permanent cell line and thus become a tumor. When abnormal cells do amass in a crypt, however, they form a polyp, which can become a large polyp if further mutations occur. Ten to twenty percent of large polyps develop into cancerous tumors.

Normally, the crypt is *homeostatic*—that is, it maintains an internal equilibrium between cell proliferation and cell loss due to death or shedding. Each cell is responsible for its own replication (which occurs when more cells are needed), as well as for its own self-destruction if



Courtesy of the National Institute of Cancer of the U.S. National Institutes of Health

FIGURE 5.11 A representation of the stages of tumor growth in the colon. Diagram courtesy of the National Institute of Cancer of the U.S. National Institutes of Health (www.cancer.gov/cancertopics/pdq/treatment/colon/Patient/page2).

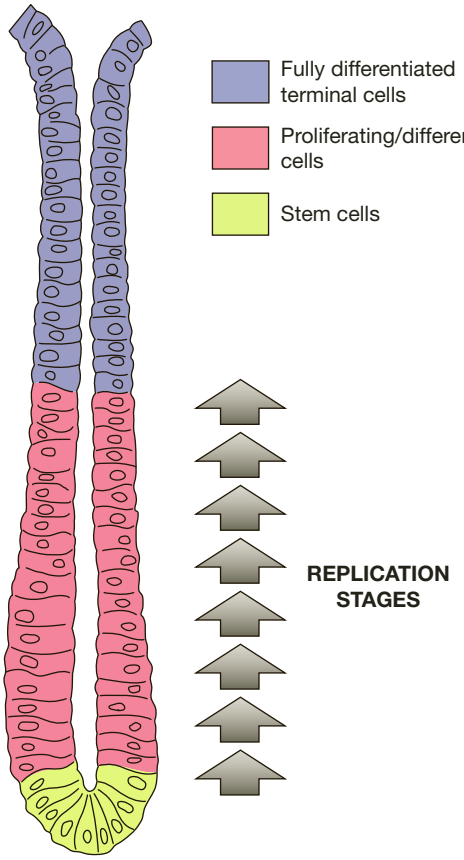


FIGURE 5.12 Stem cells anchored at the bottom of each crypt continually produce proliferating/differentiating daughter cells, which replicate an estimated eight times, resulting in a bubble of cells that rises to the epithelial surface.

the cell becomes too damaged. Thus, the initiation of a tumor requires two basic malfunctions: the failure of the cell first to control its own division and second to recognize that it is too damaged and must undergo apoptosis. Cancer cells are those that have "escaped apoptotic control" and that have bypassed all genetic mechanisms that prevent unnecessary growth. Once these control mechanisms have failed, the cell may replicate without bound.

In this section, we present several models for cell population growth in colorectal cancer developed by Matthew Johnston and colleagues at the University of Oxford. The basic model views a crypt as made up of three different types of cells. *Stem* cells reside near the bottom of the colorectal crypt; they can produce a variety of cell types required for tissue renewal and regeneration after injury. Stem cells divide to produce semi-differentiated or *transit* cells. These cells migrate up the crypt wall toward what is called the *luminal* surface. As the transit cells proceed upward, they differentiate into several types of cells. Once they reach the top of the crypt, the differentiated cells will eventually die or are shed and transported away. For simplicity, we divide the process of differentiation into three steps, stem cells (N_0), semi-differentiated cells (N_1), and fully differentiated cells (N_2).

B. Initial Model

To arrive at the equations in our first model, we need to determine the rate of change of each cellular population over time. In general, this rate should be equal to the difference between the number of cells that are added to the population and the number of cells that leave the population at the end of each cycle.

We model stem cells first. Stem cells can only die, differentiate, or renew at the end of their cycles. We assume that a certain fraction (α_1) die, a certain fraction (α_2) differentiate, and a certain fraction (α_3) renew themselves. Thus, $\alpha_3 N_0$ cells undergo renewal and $\alpha_3 N_0$ cells are added to the population (since each cell that undergoes renewal results in two renewed cells, the net addition to the population is one cell), while $\alpha_1 N_0$ cells are lost to death and $\alpha_2 N_0$ to differentiation. This analysis gives us the differential equation

$$\frac{dN_0}{dt} = \alpha_3 N_0 - \alpha_1 N_0 - \alpha_2 N_0 = (\alpha_3 - \alpha_1 - \alpha_2) N_0 \tag{34}$$

For the semi-differentiated cell population, we have a similar situation: the contribution of the renewal process is $\beta_3 N_1$ cells, while $\beta_1 N_1$ and $\beta_2 N_1$ cells die and are differentiated, respectively. We also have to take into account the stem cells that differentiate and join the population of semi-differentiated cells, so we arrive at the equation

$$\frac{dN_1}{dt} = \beta_3 N_1 - \beta_1 N_1 - \beta_2 N_1 + \alpha_2 N_0 = (\beta_3 - \beta_1 - \beta_2) N_1 + \alpha_2 N_0 \tag{35}$$

The cells that are added to the fully differentiated cells are those that differentiated from N_1, while those that leave the population are those that die. This leaves us with the equation

$$\frac{dN_2}{dt} = \beta_2 N_1 - \gamma N_2 \tag{36}$$

for the fully differentiated cellular population. Fig. 5.13 represents this model in a schematic form.

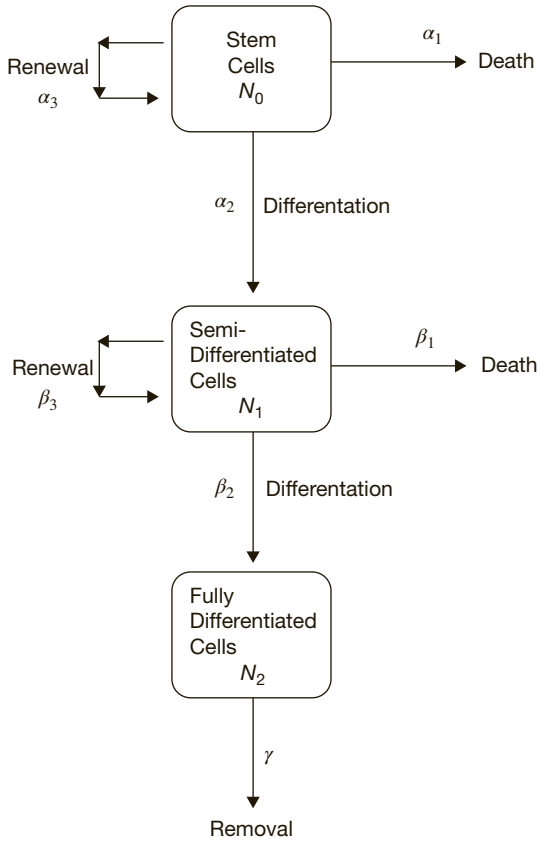


FIGURE 5.13 A schematic diagram for the initial model of colon crypt cells.

Thus, our initial model is the trio of differential equations

$$\frac{dN_0}{dt} = (\alpha_3 - \alpha_1 - \alpha_2)N_0$$

$$\frac{dN_1}{dt} = (\beta_3 - \beta_1 - \beta_2)N_1 + \alpha_2 N_0$$

$$\frac{dN_2}{dt} = \beta_2 N_1 - \gamma N_2$$

We can solve the equations of this model to find explicit representations of all three cell populations as functions of time if the αs and the βs are constants. The first equation is the differential equation for simple exponential growth with constant rate $\alpha_3 - \alpha_1 - \alpha_2$. Thus, the solution is

$$N_0(t) = N_{00}e^{(\alpha_3-\alpha_1-\alpha_2)t} = N_{00}e^{\alpha t} \tag{37}$$

where $N_{00} = N_0(0)$ and $\alpha = \alpha_3 - \alpha_1 - \alpha_2$.

To solve the second equation for the number of semi-differentiated cells $\frac{dN_1}{dt} = (\beta_3 - \beta_1 - \beta_2)N_1 + \alpha_2 N_0$, we first let $\beta = \beta_3 - \beta_1 - \beta_2$, so the equation takes the form $\frac{dN_1}{dt} = \beta N_1 + \alpha_2 N_0$ or $\frac{dN_1}{dt} - \beta N_1 = \alpha_2 N_0$. Using Eq. (37), we have

$$\frac{dN_1}{dt} - \beta N_1 = N_{00}\, e^{\alpha t} \tag{38}$$

Eq. (38) is an example of a *first-order linear differential equation*. This particular equation can be solved by first multiplying both sides by the term $e^{-\beta t}$, which is never zero. This action transforms Eq. (38) into

$$e^{-\beta t}\frac{dN_1}{dt} - \beta e^{-\beta t} N_1 = N_{00}\, e^{\alpha t} e^{-\beta t} \tag{39}$$

Observe that the left-hand side of Eq. (39) is the derivative, with respect to t, of $e^{-\beta t} N_1$ so we may rewrite this equation as

$$(\mathrm{e}^{-\beta t} N_1)' = N_{00}\, \mathrm{e}^{(\alpha-\beta)t} \tag{40}$$

We may then integrate both sides of Eq. (40) with respect to t and then multiply through by the nonzero factor $e^{\beta t}$ to obtain

$$N_1(t) = \frac{N_{00}}{\alpha - \beta} e^{\alpha t} + Ce^{\beta t} \tag{41}$$

assuming $\alpha \neq \beta$. If the initial number of semi-differentiated cells is $N_{10} = N_1(0)$, then we have

$$N_{10} = \frac{N_{00}}{\alpha - \beta} + C$$

so that

$$N_1(t) = \frac{N_{00}}{\alpha - \beta} e^{\alpha t} + \left(N_{10} - \frac{N_{00}}{\alpha - \beta}\right) e^{\beta t} \tag{42}$$

The technique for solving the differential equation for the fully differentiated cells is quite similar to the one we've just carried out. We begin by rewriting Eq. (36),

$$\frac{dN_2}{dt} = \beta_2 N_1 - \gamma N_2$$

as

$$\frac{dN_2}{dt} + \gamma N_2 = \beta_2 N_1$$

Next we multiply through by $e^{\gamma t}$ and use our solution for N_1 to obtain:

$$(e^{\gamma t} N_2)' = \frac{\beta_2 N_{00}}{\alpha - \beta} e^{\alpha t} e^{\gamma t} + \beta_2 \left(N_{10} - \frac{N_{00}}{\alpha - \beta}\right) e^{\beta t} e^{\gamma t} \tag{43}$$

Integrating Eq. (43) with respect to t and dividing through by $e^{\gamma t}$ yields

$$N_2(t) = \frac{\beta_2 N_{00}}{(\alpha-\beta)(\alpha+\gamma)} e^{\alpha t} + \frac{\beta_2\left(N_{10} - \frac{N_{00}}{\alpha-\beta}\right)}{\beta+\gamma} e^{\beta t} + Ce^{-\gamma t} \tag{44}$$

for some constant C. [Note that we are assuming that $\alpha \neq \beta$.]

Now that we have found explicit formulas for N_0, N_1, and N_2 as functions of t, let's examine the stability and long-term behavior associated with this model.

The stem cell population $N_{00}e^{\alpha t}$ grows exponentially, decays exponentially, or remains constant when $\alpha > 0$, $\alpha < 0$, or $\alpha = 0$, respectively. Since $\alpha = \alpha_3 - \alpha_1 - \alpha_2$, in the case in which $\alpha_1 + \alpha_2 + \alpha_3 = 1$, we have stability of the stem cells exactly when $0 = \alpha_3 - \alpha_1 - \alpha_2 = \alpha_3 - (1 - \alpha_3)$—that is, $\alpha_3 = ½$. Stability of the stem cells in this particular case corresponds to the situation when precisely half of them renew, since α_3 is the renewal fraction.

Let's look now at the semi-differentiated cell population when stem cells are stable $(\alpha = 0)$. We have

$$N_1(t) = \frac{N_{00}}{\alpha-\beta} e^{\alpha t} + \left(N_{10} - \frac{N_{00}}{\alpha-\beta}\right) e^{\beta t} = \frac{N_{00}}{-\beta} + \left(N_{10} + \frac{N_{00}}{\beta}\right) e^{\beta t} \tag{45}$$

If $\beta > 0$, then the semi-differentiated cells grow exponentially without bound, and if $\beta = 0$, the number of these cells remains constant at N_{10}. The third case, in which β is negative, is an interesting one, since here in the long term

$$\lim_{t\to\infty} N_1(t) = \lim_{t\to\infty} \frac{N_{00}}{-\beta} + \left(N_{10} + \frac{N_{00}}{\beta}\right) e^{\beta t} = \frac{N_{00}}{-\beta} = \frac{N_{00}}{\beta_1 + \beta_2 - \beta_3} \tag{46}$$

and

$$\lim_{t\to\infty} N_2(t) = \lim_{t\to\infty} \left[\frac{\beta_2 N_{00}}{(\alpha-\beta)(\alpha+\gamma)} e^{\alpha t} + \frac{\beta_2\left(N_{10} - \frac{N_{00}}{\alpha-\beta}\right)}{\beta+\gamma} e^{\beta t} + Ce^{-\gamma t}\right] = \frac{\beta_2 N_{00}}{(\alpha-\beta)(\alpha+\gamma)} \tag{47}$$

since $\gamma > 0$. Since $\alpha = 0$ when stem cells are stable, the limit simplifies to $-\dfrac{\beta_2 N_{00}}{\beta\gamma}$.

An interesting conclusion of this analysis is that a mutation or other alteration that changes the value of β or γ can lead to a new steady state in the size of the populations of the transit and fully differentiated cells, provided that α remains zero.

Stability for the stem cell population, however, requires that α remain zero or, equivalently, that α_3 be exactly 0.5. We call a model *structurally unstable* if the maintenance of a certain property requires that some parameter take on a particular numerical value. Since stem cells are differentiating, replicating, or renewing constantly, mutations are very likely to occur. It is unrealistic to expect that α will remain zero.

To obtain a model that exhibits structural stability, we need to build in some form of feedback that will maintain homeostasis, the ability to maintain internal equilibrium.

Johnston and his associates explored two possible feedback mechanisms: *linear* and *saturating*. Both build on the idea that the proportion of cells differentiating may depend on the sizes of the cell populations themselves.

C. Linear Feedback

In the linear feedback model, we assume that the per-capita rate of differentiation is proportional to the population. Thus, for the stem cells, we replace the original differential equation (Eq. (34)),

$$\frac{dN_0}{dt} = \alpha_3 N_0 - \alpha_1 N_0 - \alpha_2 N_0 = (\alpha_3 - \alpha_1 - \alpha_2)N_0$$

by

$$\frac{dN_0}{dt} = \alpha_3 N_0 - \alpha_1 N_0 - (\alpha_2 + k_0 N_0)N_0 = \big(\big[\alpha_3 - \alpha_1 - \alpha_2\big] - k_0 N_0\big)N_0 \tag{48}$$

for some positive constant k_0.

In this revised model, the stem cells exhibit logistic growth with a carrying capacity of $N_0^* = \dfrac{\alpha_3 - \alpha_1 - \alpha_2}{k_0} = \dfrac{\alpha}{k_0}$. If α is positive, then every solution of Eq. (48) approaches N_0^*, whereas if α is negative, the solutions approach $N_0 = 0$. There are no values of the parameters that permit unbounded growth in the stem cell population. Linear feedback for the semi-differentiated or transit cells replaces Eq. (35),

$$\frac{dN_1}{dt} = \beta_3 N_1 - \beta_1 N_1 - \beta_2 N_1 + \alpha_2 N_0 = (\beta_3 - \beta_1 - \beta_2)N_1 + \alpha_2 N_0$$

with

$$\begin{aligned}\frac{dN_1}{dt} &= \beta_3 N_1 - \beta_1 N_1 - (\beta_2 + k_1 N_1)N_1 + (\alpha_2 + k_0 N_0)N_0 \\ &= [(\beta_3 - \beta_1 - \beta_2) - k_1 N_1]N_1 + (\alpha_2 + k_0 N_0)N_0 \\ &= [\beta - k_1 N_1]N_1 + (\alpha_2 + k_0 N_0)N_0\end{aligned} \tag{49}$$

for some positive constant k_1.

Now the right-hand side of Eq. (49) is a quadratic expression in N_1 of the form $-\big[k_1 N_1^2 - \beta N_1 - C\big]$, which has value zero when $N_1 = \dfrac{\beta \pm \sqrt{\beta^2 + 4k_1 C}}{2k_1}$, where $C = (\alpha_2 + k_0 N_0)N_0$. Thus, there is one positive, stable steady state for the transit cells

$$N_1 = \frac{\beta + \sqrt{\beta^2 + 4k_1(\alpha_2 + k_0 N_0^*)N_0^*}}{2k_1} \tag{50}$$

As with the stem cells, there are no values of the parameters that result in exponential growth of the transit cells.

The differential equation for the fully differentiated cells in the linear feedback model becomes

$$\frac{dN_2}{dt} = -\gamma N_2 + (\beta_2 + k_1 N_1)N_1 \tag{51}$$

and hence their population approaches a steady state of

$$N_2^* = \frac{N_1^*(\beta_2 + k_1 N_1^*)}{\gamma} \tag{52}$$

The linear feedback model predicts that the stem cell population will sustain itself and approach a nonzero steady state provided the renewal rate α_3 exceeds a critical size $(\alpha_1 + \alpha_2)$. A mutation that alters the parameters will produce new values for the steady state populations, but unbounded growth can only result if a genetic "hit" knocks out the feedback mechanism.

D. Saturating Feedback

We turn now to the saturating feedback model. This variation of our original model also assumes that as the number of stem or transit cells increases, their rates of differentiation also grow. In contrast to the linear feedback model that posits a per capita proportion response, the saturating feedback model assumes a maximum per capita rate of differentiation.

Hence, instead of replacing α_2 with $\alpha_2 + k_0 N_0$, as we did in the linear feedback approach, we replace α_2 with $\alpha_2 + \dfrac{k_0 N_0}{1 + m_0 N_0}$ where k_0 and m_0 are positive constants. Similarly, we replace β_2 with $\beta_2 + \dfrac{k_1 N_1}{1 + m_1 N_1}$, where k_1 and m_1 are also positive constants.

Thus, the dynamic equations of the saturating feedback model are

$$\frac{dN_0}{dt} = (\alpha_3 - \alpha_1 - \alpha_2)N_0 - \frac{k_0 N_0^2}{1 + m_0 N_0} = \alpha N_0 - \frac{k_0 N_0^2}{1 + m_0 N_0} \tag{53}$$

$$\begin{aligned}\frac{dN_1}{dt} &= (\beta_3 - \beta_1 - \beta_2)N_1 - \frac{k_1 N_1^2}{1 + m_1 N_1} + \alpha_2 N_0 + \frac{k_0 N_0^2}{1 + m_0 N_0} \\ &= \beta N_1 - \frac{k_1 N_1^2}{1 + m_1 N_1} + \alpha_2 N_0 + \frac{k_0 N_0^2}{1 + m_0 N_0}\end{aligned} \tag{54}$$

$$\frac{dN_2}{dt} = -\gamma N_2 + \beta_2 N_1 + \frac{k_1 N_1^2}{1 + m_1 N_1} \tag{55}$$

For the stem cell population, we see from Eq. (53) that $\dfrac{dN_0}{dt}$ is zero when $N_0 = 0$ or $N_0 = \dfrac{\alpha}{k_0 - \alpha m_0}$. If α is negative, then $\dfrac{dN_0}{dt}$ is negative and the population of stem cells will move toward extinction, $N_0 = 0$, regardless of the initial level. For the crypt to maintain a positive stem cell population, therefore, we need α to be positive. For the second possible steady state, $N_0 = \dfrac{\alpha}{k_0 - \alpha m_0}$, to be positive we need $0 < \alpha < \dfrac{k_0}{m_0}$. In this case, $\dfrac{dN_0}{dt}$ will be

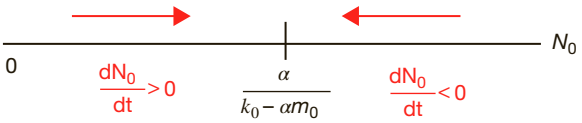


FIGURE 5.14

positive when $N_0 < \dfrac{\alpha}{k_0 - \alpha m_0}$ and negative if $N_0 > \dfrac{\alpha}{k_0 - \alpha m_0}$. The stem cell population will necessarily approach $N_0^* = \dfrac{\alpha}{k_0 - \alpha m_0}$ in the long run. Fig. 5.14 shows a schematic view.

Should it be the case that $\alpha > \dfrac{k_0}{m_0}$, then there is no positive steady state for the stem cells, and their numbers will grow unboundedly. It's possible then that a series of genetic hits increasing α by decreasing the death or differentiation rates or increasing the proliferation rate could move the crypt through increasing steady cell populations until α finally exceeds the critical $\dfrac{k_0}{m_0}$ rate, after which unbounded growth will take place.

Let's consider the case $0 < \alpha < \dfrac{k_0}{m_0}$, in which we have a steady state for the stem cell population, N_0^*. We investigate the long-term behavior of the transit cell population N_1 that satisfies the differential equation

$$\begin{aligned}\frac{dN_1}{dt} &= \beta N_1 - \frac{k_1 N_1^2}{1 + m_1 N_1} + \alpha_2 N_0 + \frac{k_0 N_0^2}{1 + m_0 N_0} \\ &= N_1 \left(\beta - \frac{k_1 N_1}{1 + m_1 N_1} \right) + \alpha_2 N_0 + \frac{k_0 N_0^2}{1 + m_0 N_0} \\ &= N_1 \left(\beta - \frac{k_1}{\dfrac{1}{N_1} + m_1} \right) + \alpha_2 N_0 + \frac{k_0 N_0^2}{1 + m_0 N_0}\end{aligned} \tag{56}$$

where $\beta = \beta_3 - \beta_1 - \beta_2$. The last two terms will approach a fixed positive constant $D = \alpha_2 N_0^* + \dfrac{k_0 N_0^{*2}}{1 + m_0 N_0^*}$. If β exceeds the threshold value of $\dfrac{k_1}{m_1}$, then $\left(\beta - \dfrac{k_1}{\frac{1}{N_1} + m_1} \right)$ will be positive for all positive values of N_1. Thus, when $\beta > \dfrac{k_1}{m_1}$, the derivative dN_1/dt remains positive and is bounded away from zero. Hence, the transit cells will undergo unbounded growth.

If β is below the threshold value $\dfrac{k_1}{m_1}$, we can determine a steady state for the transit cell population by solving $\dfrac{dN_1}{dt} = 0$, where $\dfrac{dN_1}{dt} = \beta N_1 - \dfrac{k_1 N_1^2}{1 + m_1 N_1} + D$. This equation becomes

$$\beta N_1 (1 + m_1 N_1) - k_1 N_1^2 + D(1 + m_1 N_1) = 0$$

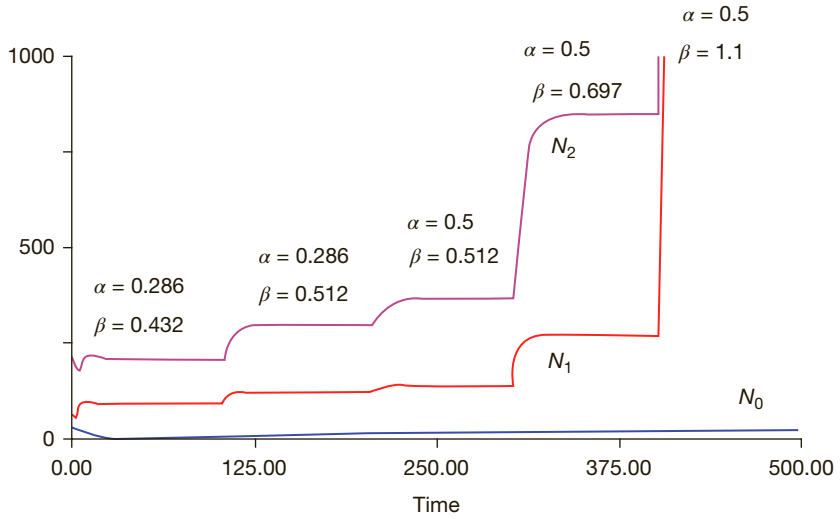


FIGURE 5.15 Cell population growth affected by "genetic hits."

Multiplying out and collecting like terms yields the quadratic equation

$$(k_1 - \beta m_1)N_1^2 - (\beta + Dm_1)N_1 - D = 0 \tag{57}$$

which has a single positive root

$$N_1^* = \frac{\beta + Dm_1 + \sqrt{(\beta - dm_1)^2 + 4Dk_1}}{2(k_1 - \beta m_1)}.$$

Observe also that if N_1^* is the steady state for the transit cells, then the steady state for the fully differentiated cells is $N_2^* = \dfrac{\beta_2 N_1^* + \frac{k_1 N_1^{*2}}{1 + m_1 N_1^*}}{\gamma}$.

The saturating feedback model thus shows that successive mutations increasing β may also give rise to a sequence of increasing steady state cell populations, which become unbounded only if β a surpasses a critical threshold value.

We can illustrate this process with an example. Suppose k_0 and m_0 are each 0.1 and k_1 and m_1 are each 0.01 while α_2 and β_2 are each 0.3 and $\gamma = .323$. The critical threshold values are $k_0/m_0 = 1$ and $k_1/m_1 = 1$. We assume that initially $\alpha = 0.286$ and $\beta = 0.432$. In the long term, the cell populations will approach $N_0^* = 4$, $N_1^* = 85$, and $N_2^* = 200$. Now we examine what happens if there are a series of genetics hits, equally spaced in time.

At time 100, assume that there is a mutation causing β to increase to 0.512. This change does not affect N_0^*, but makes $N_1^* = 114$ and $N_2^* = 294$. Then, at time 200, another genetic hit raises α to 0.5; the new steady states are $N_0^* = 10$, $N_1^* = 134$, and $N_2^* = 361$. A third mutation, occurring at time 300, bumps β up to 0.697, yielding $N_0^* = 10$, $N_1^* = 266$, and $N_2^* = 847$. Finally, a fourth mutation at time 400 pushes β above the critical threshold level to 1.1. Then both the semi-differentiated and fully differentiated cell populations grow exponentially. Fig. 5.15 shows the graphs of the cell populations over time under this scenario; these were obtained using the Euler approximation technique for differential equations introduced in Chapter 2.

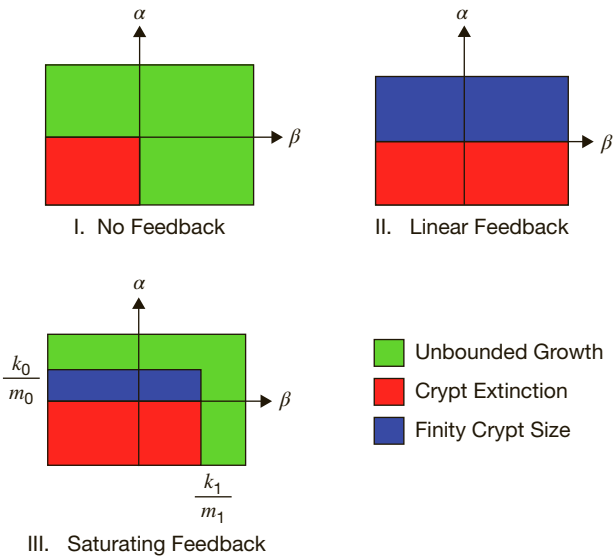


FIGURE 5.16 Plots of the regions of stability of the cell population models.

Fig. 5.16 provides a visual summary of our stability findings. In the first model (Eqs. (34−36)) with no feedback, the only stable solutions lie on the line $\alpha = 0, \beta < 0$. Otherwise, there is either unbounded growth if either α or β is positive or extinction if both α and β are negative. In the linear feedback model, there is a steady state whenever $\alpha > 0$ and extinction when $\alpha < 0$. There is never unbounded growth.

In the saturating feedback model, steady states are possible only in the strip $0 < \alpha < \frac{k_0}{m_0}$, $\beta < \frac{k_1}{m_1}$. If the parameters lie outside this strip, there is either extinction or unbounded growth.

The feedback models created by Johnston et al. predict results fully consistent with real-world observations. As they note,

> *This process simulates the widely assumed multistage process of carcinogenesis. Successive mutations could cause parameter changes which incrementally raise the size of the steady state. However, once the mutations have accumulated to a certain degree, and the parameters are raised above a certain threshold, unregulated cell population growth occurs and the tumor grows exponentially.*

V. Historical and Biographical Notes

A. Benjamin Gompertz

Benjamin Gompertz was born in London on March 5, 1779, into a distinguished Jewish merchant family that had emigrated to England from Holland. Gompertz displayed brilliance from his boyhood and had a deep thirst for knowledge; when his parents removed the candles to prevent him from injuring his health by studying too late into the night, Gompertz would frequently continue his reading by moonlight in the garden.

Until well into the 19th century, British universities required oaths of allegiance to the Protestant Church of English, effectively barring Jews from higher education. Blocked

from obtaining a formal college education by religious discrimination, Gompertz was largely self-taught. At the age of 18, he joined the Spitalfields Mathematical Society, which later merged with the Royal Astronomical Society of London. Gompertz contributed regularly to the *Gentleman's Mathematical Companion*, winning the journal's prize competition every year from 1812 to 1822. Later work on complex variables, published under the title *The Principles and Application of Imaginary Quantities*, established his reputation as an eminent mathematician, earning him election as a Fellow of the Royal Society in 1819.

In 1821, Gompertz was an unsuccessful candidate for the position of actuary at the newly established Guardian Insurance Office. Partly in response to a report that Gompertz had been denied the job because of his religion, but also to take advantage of Gompertz's mathematical abilities, his brother-in-law, Sir Moses Montefiore, and Nathan Rothschild set up the Alliance British and Foreign Life Assurance Company. Gompertz served as the very successful business's actuary and chief executive officer.

Public domain

BENJAMIN GOMPERTZ

The profit of a company selling life insurance policies lies in collecting more in premiums from customers than it has to pay out in death benefits in any given year. To determine how much to charge in premiums, the company needs accurate estimates of how many of its policyholders are likely to die in the next year. Toward this end, actuaries and statisticians had compiled "life tables" that recorded how many people had died in a community, along with the age and cause of death. Before Gompertz, a life table served mainly as a way to compute the number of persons surviving to a later age out of a given number alive at an earlier age. Gompertz sought to discover the laws that produced consistent age patterns of death. The law of human mortality associated with his name was

propounded in papers published in the *Philosophical Transactions* in 1820 and 1825, with a supplementary paper published there in 1862. Gompertz's contributions ushered in a new era for actuarial science. His insights have remained central to the study of human mortality.

What Gompertz observed in the early 1820s was an exponential rise of death rates in a population between sexual maturity and old age. He attributed this phenomenon to an underlying law of mortality: "the average exhaustions of a man's power to avoid death" are such that "at the end of equal infinitely small intervals of time" he loses "equal portions of his remaining power to avoid destruction." Well versed in the ideas and tools of calculus (he was once described as "the last of the learned Newtonians"), Gompertz quickly translated this verbal description into the differential equation (11).

Physical ill health forced Gompertz's retirement from active work in 1848, but he continued to work on problems in mathematics and astronomy. In 1850, he published a sequel, *Hints on Porisms*, to his earlier books on complex numbers. He also wrote on comets and meteors, contributed a paper on human mortality to the International Statistical Congress in 1860, and was working on a paper for the London Mathematical Society at the time of his death on July 14, 1865.

Although Gompertz made many significant contributions to astronomy and the actuarial sciences, he was equally enthralled by the interplay of mathematical ideas detached from practical applications. Consider, for example, these thoughts of his:

> *In the contemplation of the sciences there is, besides the pleasure arising from the acquirement of knowledge of practical utility, a peculiar charm bestowed by the reasoning faculty in a well-directed pursuit of facts; and though the results shown by the arguments are frequently considered to be the only objects of value by the unlearned, the man of absolute scientific ardour will often, whilst he is enraptured with the argument, have not the least interest for the object for which his argument was instituted.*

The renowned American geneticist Sewall Wright (1889–1988) was apparently the first to suggest using the Gompertz curve to model biological growth. In a 1926 review of Raymond Pearl's *The Biology of Population Growth*, Wright pointed out that Pearl had focused on one particular S-curve, the logistic, but that other mathematical forms might better reflect nature. Wright contrasted the growth of *populations* to that of individual *organisms*:

> *Populations of fruit flies and certain human populations, notably the United States, follow this simple [logistic] law of growth very satisfactorily. Indeed, populations from yeast cells to man appear to conform to it much more closely than do individual organisms, which in general show a point of inflection at an earlier stage. Perhaps this points to a fundamental difference in the nature of the limiting conditions. In populations the inherent reproductive capacity of the individuals is not necessarily changed as the cycle advances. The restraint to growth is external and apparently does increase as a certain limit is approached in some conformity with the simple rule described above. In organisms, on the other hand, the damping off of growth depends more on internal changes in the cells themselves, the process which Minot called cytomorphosis. The average growth power as measured by the percentage rate of increase tends to fall at a more or less uniform percentage rate, leading to asymmetrical types of S-shaped curves of which the form* $\log \log k/y = a(b-x)$ *is a simple example, instead of the logistic curve,*

$$\log\left(\frac{k}{y} - 1\right) = a(b - x)$$

In 1964, Anna Kane Laird ("Dynamics of tumor growth," *British Journal of Cancer* **18**: 490−502) was the first scientist to use the Gompertz curve to fit data of growth of tumors successfully.

B. Ludwig von Bertalanffy

Ludwig von Bertalanffy, in the words of his biographer Mark Davidson, "may well be the least well known titan of the Twentieth Century. As the father of the interdisciplinary school of thought known as general systems theory, he made important contributions to biology, medicine, psychiatry, psychology, sociology, history, education, and philosophy. Yet he spent his life in semi-obscurity and he survives today mostly in footnotes."

Born in a little village near Vienna on September 19, 1901, Bertalanffy began his studies with history of art and philosophy at the University of Innsbruck and then at the University of Vienna. He completed his doctoral thesis on the German physicist and philosopher Gustav Theodor Fechner in 1926, and published his first book on theoretical biology, *Modern Theories of Development*, 2 years later.

Kenneth Boulding, a renowned economist and an early follower of Bertalanffy in the creation of general systems theory, provides the following picture of Bertalanffy:

> *A man I remember as being like no other—kindly, rather shy, a curious mixture of confidence that he was saying something important and diffidence that grew out of lack of people to receive it. He presented a façade that was almost a caricature of the Viennese professor, but behind the façade one felt a remarkable mind and spirit with an extraordinary sense both of the immense complexity of the real world and strong faith that it was not wholly beyond our grasp.*

Bertalanffy held a number of teaching and research positions during his career, the longest at the University of Vienna from 1934 to 1948. Awarded a Rockefeller Foundation fellowship in 1937, he spend a year in the United States, studying developments there in biology and giving seminars on philosophy and science. Ludwig and his wife Maria hoped to remain in America to avoid the gathering storm of anti-Semitism and German militarism, but he was unable to secure the promise of a permanent position and reluctantly returned to Europe. During World War II, von Bertalanffy conducted pioneering research on cancer and lectured on biology to classes of hundreds of medical students.

The Third Reich considered Bertalanffy somewhat of a subversive. He had published an essay "The Science of Life and Education" in 1930 in which he denounced biological theories that were used to justify racism. Nazi book burners destroyed copies of this work. He often provoked arguments with Nazi sympathizers on the faculty, and the Bertalanffys had a number of close friends who were Jewish. During the Russian siege of Vienna in April 1945, his family was displaced from his house. "After the siege," Davidson reports, "when they made their way back home through rubble and corpses, they found a smouldering wreckage where their apartment building had stood. Their neighborhood had been destroyed by German flame throwers as part of the Nazi policy of scorched-earth retreat." All their personal possessions were lost, including a library of 15,000 books and major parts of books Bertalanffy had been writing. At the university, he discovered that a bomb had destroyed his office and lab, and that his entire department had been ransacked and was in shambles.

He also worked at the University of London, the University of Montreal, the University of Ottawa, the Center for Advanced Study in the Behavioral Sciences, Mount Sinai Hospital of Los Angeles, the Menninger Foundation, the University of Alberta, and the State University of New York. He died from a sudden heart attack on June 12, 1972, in Buffalo, New York.

Bertalanffy was the first scientist to undertake a mathematically rigorous approach to the understanding of biochemical synergies. The interrelationships of individual cells, individual organisms, individual people and societies intrigued Bertalanffy, who recognized that studying these interdependencies and interactions was critical to appreciating and comprehending living things. As he phrased it

> *Entities of an essentially new sort are entering the sphere of scientific thought. Classical science in its diverse disciplines, be it chemistry, biology, psychology or the social sciences, tried to isolate the elements of the observed universe—chemical compounds and enzymes, cells, elementary sensations, freely competing individuals, what not—expecting that, by putting them together again, conceptually or experimentally, the whole or system—cell, mind, society—would result and be intelligible. Now we have learned that for an understanding not only the elements but their interrelations as well are required . . . .*

Public domain

Ludwig von Bertalanffy

Bertalanffy lived through times that saw the evil consequences of Hitlerism, Stalinism, McCarthyism, jingoism, and chauvinism, yet he maintained a single standard of morality toward all people. For him, Davidson observes, "a wrongful act was equally wrong whether perpetrated by capitalist or communist, archbishop or atheist, professor or pipefitter, a friend or foe."

Although he made significant advances in many disciplines, Bertalanffy maintained a good degree of modesty about his achievements. Shortly before his death, he commented to his students

> *I cannot offer a wonder drug or panacea for the salvation of society. I'm not Mister-Know-It-All and I cannot promise you answers. But perhaps our discussion will help us a little bit to better understand those pressing problems with which we are confronted. The only thing we can hope for is perhaps getting a little bit wiser about our problems and about what can be done. So let us discuss these things as far as you and I are able to do so—because you see, I want to be of some use to you. If I succeed in making a very tiny contribution in that way, then I would be satisfied."*

C. Anna Kane Laird

Born in New York City on June 7, 1922, Anna Kane Laird was a distinguished biologist and psychiatrist. She earned her B.A. and M.D. degrees at the University of Pennsylvania in the 1940s and finished a doctoral degree at the University of Wisconsin in 1952. Dr. Laird was a postdoctoral fellow in cancer research at the U.S. Public Health Service and the American Cancer Society before she assumed a long-term position as biologist and pathologist at the Argonne National Laboratory.

Photograph reproduced by permission of the University of Pennsylvania Archives

Anna Kane Laird near the time of her graduation from the University of Pennsylvania Medical School in 1946.

Laird published dozens of papers and monographs, many of them co-authored with her husband Ambrose Donald Barton. Much of her work centered on the mathematical analysis of animal growth, biological time, heritable factors in normal growth, and properties of malignant growth.

She also completed residencies in clinical pathology and psychiatry at the University of Wisconsin. In the latter part of her career, she was a psychiatrist in Madison, Wisconsin. Laird died on June 10, 2007.

Laird pioneered in the use of the Gompertz model in biology. She saw that real-world data on tumor growth was not consistent with the prevailing model at the time. As she wrote in a groundbreaking paper "Dynamics of Tumor Growth,"

> *It is commonly believed that tumor growth under ideal conditions is a simple exponential process terminated by the exhaustion of the nutritional support provided by the host. However, a survey of the literature shows that exponential growth of tumors has been observed only rarely and then only for relatively brief periods. When we consider those tumors whose growth has been followed over a sufficiently extensive range (100 to 1000-fold range of growth or more), we find that nearly all such tumors grow more and more slowly as the tumor gets larger, with no appreciable period of growth at a constant specific growth rate as would be expected for simple exponential growth.*

Laird described the experimental observations as demonstrating "continuous deceleration of growth." She viewed the dynamics of tumor growth as one governed by an equation of the form $dV/dt = rV$ where r, instead of being a constant, is itself an exponentially decreasing function of time—that is, $dr/dt = -kr$. We saw this system of differential equations earlier (Eq. (19)) and showed that it is equivalent to the Gompertz approach.

D. Sylvanus Alexander Tyler Sr.

Laird was assisted in the mathematical analysis that appeared in this paper by Sylvanus Alexander Tyler Sr. Tyler (August 21, 1914−July 23, 1986) was a notable African American mathematician and biostatistician. He received a bachelor's degree from Fisk University and a master's degree from the University of Chicago, where he wrote his thesis, *The Projective Generation of Curves*. After service in the Signal Corps in World War II, Tyler began a 34-year career with the Argonne National Laboratory in Chicago, where Laird and Barton also worked. He published more than 70 articles and books. Tyler also compiled statistics for a 1945 volume *Black Metropolis*: *A Study of Negro Life in a Northern City*.

Tyler was co-author with Laird and Barton of a 1965 paper "Dynamics of Normal Growth," which demonstrated that a modified Gompertz function fits growth of the normal mammalian organism, from early embryonic periods to young adulthood.

Photograph reproduced by permission of Sylvanus Tyler Jr.

Sylvanus Tyler Sr.

Laird's use of the Gompertz model inspired Lawrence Norton and Richard Simon to propose a new treatment regime of cancerous tumors. Norton had a patient with Hodgkin's disease who had a spectacular response to chemotherapy. He stayed in complete remission for about a year, then his tumor recurred in the same location with the same pathology. After subsequent treatment. The patient went back into complete remission. "But something puzzled me," Dr. Norton said. "How could such a large mass disappear into complete remission, and then recur? . . . So I looked at the math, trying to understand the exponential mathematical model we used to describe tumor growth. I discovered that the mathematical model we used to tailor our treatment didn't make sense."

Norton soon found Laird's paper and confirmed that measurements of tumor growth he had collected matched the Gompertz function. Since the Gompertz curve shows that the growth rate is nearly exponential at early stages of development and slower at later stages, Norton and Simon proposed that tumors be given less time to regrow between treatments. They noted that many drugs killed cancer cells at rates proportional to tumor growth rates, so smaller tumors should be easier to eradicate than larger ones. As Charles Schmidt reported in Journal of the National Cancer Institute,

> *The Norton-Simon hypothesis flew in the face of conventional views, which held that tumor growth is exponential and that chemotherapy kills in log intervals, meaning it kills constant fractions of tumor. When it was published in the 1970s, the hypothesis was met with such fierce hostility that Norton considered leaving oncology altogether.*

Norton was able to oversee clinical trials that yielded results consistent with his hypothesis. Standard therapy changed with intervals between chemotherapy treatments were shortened, improving survival rates among patients. In 2004, the American Society of Clinical Oncology awarded its highest honor to Norton—"sweet vindication," Schmidt wrote, "for Norton and his efforts to advanced an unpopular and poorly funded topic: the application of mathematical concepts to cancer biology."

E. Matthew Johnston

Matthew David Johnston was born on April 21, 1981, in Wimbledon, England. He grew up in Ewell, a small village in Surrey, just south of London. His interest in mathematics was sparked by teachers in his primary school, Lynton Preparatory, which was operated by two sisters and their husbands for more than half a century.

Johnston finished his secondary school program at King's College School in Wimbledon. He did his undergraduate work at Trinity College Oxford, where he earned First Class Honours and won several prizes in mathematics. For 2 years, he worked evenings and weekends as a gas station attendant. "It brought in much needed funds that supported me through undergraduate study," Johnston recalls, "although it did curtail my social life a bit at the time!"

An avid tennis player in high school and college, Johnston won several singles and doubles tournaments. At Oxford, he was captain of the Trinity tennis team. In a gap year after his undergraduate degree, Johnston did voluntary duty at an Oxfam charity shop and also worked on models of road networks at WS Atkins engineering firm. He also worked at the London School of Pharmacy for 4 months on a project sponsored by Cancer Research UK, using a C++ model to analyze DNA sequencing in the human genome to ascertain the prevalence of G-quadruplex sequences that could be used as potential therapeutic targets.

He completed his doctoral degree at the University of Oxford in 2008, where he worked in the applied mathematics departments of the Oxford Centre for Industrial and Applied Mathematics and the Centre for Mathematical Biology. As a graduate student, Johnston was a leading member of a team developing mathematical models of the colonic crypt in colorectal cancer. They created discrete, continuous age-structured, and spatial models to recreate observations of a healthy crypt, and extended these models to incorporate cancerous growth.

Photo reproduced by permission of Matthew Johnston

Matthew Johnston

EXERCISES

II. A General Tumor Growth Model

1. Show that the general growth model of Eq. (2) reduces to exponential growth if $\alpha = \beta = 1$.

2. What can you say about behavior of the growth model of Eq. (2) in the general case that α and β are equal?

3. Use the fact that the volume V of a sphere is $4/3\pi r^3$ and its surface area A is $4\pi r^2$ to show that A is a constant multiple of $V^{2/3}$.

4. If an object is in the shape of cube, show that the surface area is proportional to the two-thirds power of the volume. Is the conclusion true if the shape is a circular cylinder or a cone?

5. Show that the "carrying capacity" of the Bertalanffy model of Eq. (5) is $\left(\frac{a}{b}\right)^3$.

6. Carry out the details of solving the Bertalanffy Eq. (5) to derive Eq. (6).

7. Compare/contrast direction fields for logistic and Bertalanffy equations.

III. The Gompertz Curve

8. Why does l'Hôpital's rule apply to $\lim_{x\to 0}\left[\frac{V^x - 1}{x}\right]$, and precisely why is this limit equal to $\lim_{x\to 0}\frac{V^x \ln V - 0}{1}$?

9. Show that Eq. (12) implies that for positive values of V, V is increasing whenever V is less than $e^{a/b}$ and decreasing for $V > e^{a/b}$. Can you conclude that $e^{a/b}$ is a stable equilibrium value for V?

10. Use the Gompertz equation in the form $\frac{dV}{dt} = (a - b\ln V)V$ to explain why $\lim_{t\to\infty} V(t) = e^{\frac{a}{b}}$.

11. Sketch the graph of dV/dt versus V for $0 \le V \le e^{\frac{a}{b}}$. What can you conclude about the behavior of V from this graph?

12. In solving the Gompertz differential equation, we assumed that $\int \frac{1}{u}du = \ln u + C$ rather than the more formally correct answer $\int \frac{1}{u}du = \ln|u| + C$. Were we safe in ignoring the absolute value signs?

13. Here's a different approach to solving the Gompertz differential equation $\frac{dV}{dt} = (a - b\ln V)V$.

(a) Show that the change of variable $Q = e^{-a/b}V$ transforms the Gompertz equation into the differential equation $\frac{dQ}{dt} = -bQ\ln Q$. Note that as V ranges from 0 to $e^{\frac{a}{b}}$, the variable Q will range from 0 to 1.

(b) Solve the transformed differential equation by separation of variables and integration, noting that $\frac{d}{dt}(\ln\ln Q) = \frac{1}{Q\ln Q}$.

14. Show that the Gompertz curve has a single point of inflection at the time when $\ln V = \frac{a}{b} - 1$.

15. Verify the equivalence of two different forms of the Gompertz model by substituting $b = k$ and $a = r_0 + k\ln V_0$ into the first form.

16. Use the principle of mathematical induction to prove

(a) $\sum_{i=1}^{N} i = \frac{N(N+1)}{2}$

(b) $\sum_{i=1}^{N} i^2 = \frac{N(N+1)(2N+1)}{6}$

17. With the estimated parameter values for the Gompertz model of chicken growth, show that the predicted long-range limit to the size is 4.476.

18. Carry out the details of fitting the logistic model to the chicken weight data to show that $d = 3.155450907$ and $a = 0.4124054532$.

19. Another technique for approximating V_0 is based on estimating V_∞, the upper limit of V as t gets large. Verify that $\lim_{t\to\infty} V(t) = \lim_{t\to\infty} V_0 e^{\frac{r_0}{k}} e^{\frac{r_0}{k}e^{-kt}} = V_0 e^{\frac{r_0}{k}}$ so that $V_\infty = V_0 e^{\frac{r_0}{k}}$ and $V_0 = V_\infty e^{-\frac{r_0}{k}}$. Show that for our chicken example, if we estimate V_∞ as 4.6, then $V_0 = 0.0585934$. If we use this value for V_0, show that the sum of squared errors is reduced to 0.0120392.

20. What can you conclude about the qualitative behavior of a growth model of the form $dx/dt = g(x)$ where g is concave down, increases to a unique positive maximum, and then decreases to zero? Is the solution necessarily an S-shaped curve?

21. Show that the solutions of $dx/dt = \sin x$ and $dx/dt = \sin^2 x$ on $0 \le x \le \pi$ are S-shaped curves.

22. Show that the solutions of $dx/dt = \cos x$ and $dx/dt = \cos^2 x$ on $-\pi/2 \le x \le \pi/2$ are S-shaped curves.

23. Without explicitly solving the differential equation, show that the solution of the *generalized logistic equation* $\frac{dx}{dt} = a(x-b_1)^c(b_2-x)^d$ will be an S-curve if a, b_1, b_2, c, d are positive and $b_2 > b_1$.

24. Without explicitly solving the differential equation, show that the solution of the *generalized Gompertz equation*

$$\frac{dx}{dt} = \begin{cases} a(x-b_1)^c\left(\ln\dfrac{b_2-x}{x-b_1}\right)^2 & \text{if } x > b_1 \\ 0 & \text{if } x = b_1 \end{cases}$$

will be an S-curve if a, b_1, b_2, c, d are positive and $b_2 > b_1$.

25. Without explicitly solving the differential equation, show that the solution of the *exponential sigmoid equation* $\frac{dx}{dt} = a(x-b_1)^c(e^{b_2}-e^x)^d$ will be an S-curve if a, b_1, b_2, c, d are positive and $b_2 > b_1$.

IV. Colorectal Cancer

26. Since $\alpha_1 + \alpha_2 + \alpha_3 = 1$, show that equation (34) can be rewritten as

$$\frac{dN_0}{dt} = (1 - 2\alpha_1 - 2\alpha_2)N_0$$

27. Show that Eq. (35) can be rewritten as

$$\frac{dN_1}{dt} = \beta_3 N_1 - \beta_1 N_1 - \beta_2 N_1 + \alpha_2$$
$$N_0 = (1 - 2\beta_1 - 2\beta_2)N_1 + \alpha_2 N_0$$

28. The general first-order linear differential equation, of which Eq. (38) is an example, has the form $\frac{dy}{dt} - p(t)y = q(t)$, where p and q are continuous functions of t.

(a) Show that multiplying through this equation by the *integrating factor* $e^{-\int p(t)dt}$ produces the equation $e^{-\int p(t)dt}\frac{dy}{dt} - p(t)e^{-\int p(t)dt}y = q(t)e^{-\int p(t)dt}$ where the left-hand side is exactly the derivative, with respect to t, of $e^{-\int p(t)dt}y$.

(b) Deduce from (a), that the solution of the general first-order linear differential equation has the form $y = e^{\int p(t)dt}\int q(t)e^{-p(t)dt}dt + Ce^{\int p(t)dt}$ where C is a constant whose value depends on the initial value of y.

29. Use the technique outlined in Exercise 28 to solve the differential equation

$$\frac{dy}{dt} + \tan t\, y = \cos^2 t$$

where $y(0) = 2$.

30. Use the technique outlined in Exercise 28 to solve the differential equation

$$\frac{dy}{dt} = \frac{2t}{1+t^2}y + \frac{2}{1+t^2} \quad \text{with} \quad y(0) = 2/5$$

31. Find the value of the constant C in Eq. (44) if $N_2(0) = N_{20}$.

32. Note that Eq. (46) is undefined if $\beta = 0$. What can you conclude about the population of semi-differentiated cells when $\beta = 0$?

33. Use Eq. (48) to show that under the linear feedback model, the number of stem cells N_0 will increase if $N_0 < \frac{\alpha}{k_0}$ and increase if N_0 exceeds $\frac{\alpha}{k_0}$.

34. For the saturating feedback model, verify the claim that if $0 < \alpha < \frac{k_0}{m_0}$, then $\frac{dN_0}{dt}$ will be positive when $N_0 < \frac{\alpha}{k_0 - \alpha m_0}$ and negative if $N_0 > \frac{\alpha}{k_0 - \alpha m_0}$.

35. Why is it that the quadratic in Eq. (57) has exactly one positive root?

36. Verify the claim that for the saturating feedback model, if N_1^* is the steady state for the transit cells, then the steady state for the fully differentiated cells is $N_2^* = \dfrac{\beta_2 N_1^* + \frac{k_1 N_1^{*2}}{1 + m_1 N_1^*}}{\gamma}$

SUGGESTED PROJECTS

1. Investigate the discrete form of the Gompertz model using the form $Q_{i+1} - Q_i = -bQ\ln Q_i$. Are there values of b for which chaos is observed? See our discussion of the discrete logistic model in Chapter 3. You may also wish to examine the paper by Daisuje Satoh (2000).

2. Examine methods for obtaining exact or approximate solution for the generalized Bertalanffy and Gompertz models. See Miljenko Marušić and Zeljko Bajzer (1993).

3. Michael Savageau derived the generalized Bertalanffy equation from a different set of first principles. Investigate his approach. See Savageau (1979).

4. Steven Piantadosi (1985) proposed a model of tumor growth, which views the tumor as a cell population of size N divided into a reproducing group P and a quiescent group Q. A proportion g of the Q cells reenters the reproducing group while a different proportion w die off. The P population dies off at the same rate w, but also gains in size through splitting at a constant rate a. These assumptions lead to differential equations $Q' = -gQ - wQ$ and $P' = gQ + aP - wP$. Show that you can combine these into a single differential equation $N' = aFN - wN$, where $F = P/N$ is the growth fraction. Piantadosi proposed one relationship between F and tumor volume V, while other scientists (see Cox (1980) and Matusic (1991)) considered others. Investigate the mathematical consequences of these models, and the ways that their predicted values match observed data.

5. One way to describe dynamic processes that give rise to S-curves is that they are solutions to differential equations of the form $dx/dt = g(x)$ where there are positive numbers $a < b < c$ such that

(a) $g(x) > 0$ on the interval (a, c)

(b) $g(a) = g(c) = 0$

(c) $g'(x) > 0$ on the interval (a, b)

(d) $g'(x) < 0$ on the interval (b, c)

Show that the logistic and Gompertz models are of this form for appropriate choices of g. What general properties (e.g., must there be a single point of inflection?) can you derive from conditions (a)−(d)?

6. We have presented a discussion of the *continuous* models of colon cancer dynamics investigated by Johnston and his colleagues. Their papers (2006, 2007) also develop a *discrete* model you may be interested in investigating.

7. Johnston's models assume that a stem cell can either divide to create two new stem cells or two differentiated cells. Some biologists conjecture that a single stem cell may also evolve into one stem cell and one differentiated cell. Extend Johnston's models to include this third possibility.

8. Sales over time of consumer electronic products often follow a Gompertz curve with moderate initial interest followed by near exponential growth tapering off as the market reaches saturation. Sanjay Singh (2007) found that mobile phone sales in India were accurately modeled as Gompertz growth. Investigate other markets for cell phones and other products (for example, large screen flat TV screens) whose sales curves are S-shaped to see how closely a Gompertz or a logistic model fits the available data.

You can find a listing of references and suggestions for additional reading on the book's website, www.wiley.com/college/olinick

Social Choice and Voting Procedures

CHAPTER 6

> The general will is always right and tends to the public advantage; but it does not follow that the deliberations of the people are always equally correct. Our will is always for our own good, but we do not always see what that is; the people is never corrupted, but it is often deceived.
>
> —Jean-Jacques Rousseau

I. Three Voting Situations

This chapter illustrates the use of axiomatic models by investigating some of the procedures groups of voters use to determine collective judgments from individual preferences. These procedures characteristically have certain injustices associated with them. An axiomatic approach reveals that attempts to redesign the procedures or invent new ones to avoid these inequities are doomed to frustration.

An illustrative real-world example is the U.S. Senate and its attempts to reach agreement on certain types of important issues. The model we develop concerns certain kinds of collective judgments, which are exemplified by the following three illustrations.

Example 1

The president nominates a South Carolina lawyer for a position on the U.S. Supreme Court. The Senate must decide whether to confirm the nomination or not.

Example 2

Three proposals for dealing with the dependents' deduction feature of the federal income tax have been offered. Proposal *A* calls for a substantial increase in the amount of the deduction so that it will more accurately reflect the costs of rearing children in today's economy. Proposal *B* seeks the abolition of the dependents' allowance; its advocates wish to discourage parents from planning large families. Proposal *C* is simply that the present level of the deduction be retained. The Senate must adopt one of these mutually exclusive proposals. This is an example of a *Social Choice Problem*.

Example 3

A commission on national goals asks the Senate for its evaluation of the order of importance of three current problems: the economy, the plight of urban areas, and the protection of the environment. In this situation, the Senate must indicate an *ordering* of three alternatives. This is an example of a *Social Welfare Problem*.

There are 100 senators, two from each state. Assume that there has been sufficient discussion and debate on the matters before the Senate so that each member has already determined his or her own personal preferences among the alternatives open. What procedure should be used in passing from this set of 100 individual preferences to a collective preference?

II. Two Voting Mechanisms

A. Simple Majority Voting

The first situation, that of confirming a nomination to the Supreme Court, poses little difficulty. Each senator announces a vote for or against the nominee. If a simple majority of those voting favors the nominee, the nominee is confirmed. Otherwise, the nominee is rejected. (Should the Senate split evenly, then the ballot of the vice president is counted to Vice President determine the majority position.)

Decision making by simple majority voting is, of course, the most familiar scheme for determining the collective judgment of a group of individual voters. Together with the concept that each individual has but a single vote, it forms the heart of what many would define as *democracy*. "The very essence of democratic government," wrote Alexis De Tocqueville in *Democracy in America* (1835), "consists in the absolute sovereignty of the majority; for there is nothing in democratic states which is capable of resisting it." In his first inaugural address, Abraham Lincoln observed, "Unanimity is impossible; the rule of a minority, as a permanent arrangement, is wholly inadmissible; so that, rejecting the majority principle, anarchy or despotism in some form is all that is left."

As you will see, simple majority voting is a fair and effective procedure to adopt when a group must decide between *two* alternatives or candidates. But there are many situations in which a society needs to make a choice among three or more alternatives. In each of the U.S. presidential elections of 1980, 1992, 1996, 2000, and 2004, there were strong "third-party" candidates who appealed to millions of voters as more attractive than the major party nominees. There are numerous elections for governors, state and county offices, and legislative positions where no single candidate is the first choice of a majority of the voters.

What happens, then, when the group must choose among three or more alternatives? How does the Senate actually reach a decision when faced with a situation like that described in Example 2? It adopts a procedure used by many legislative bodies: change the format of the problem from one involving a choice among three alternatives to a series of choices between two alternatives.

To illustrate this process with the income tax example, the Senate might first decide between proposals A and B on a simple majority vote. The winning proposal would then be pitted against C and the eventual winner then decided by a simple majority vote between these two alternatives.

The idea is to use the simple majority principle—because of a strong belief in its fairness—even when it may not be immediately applicable. Is there anything wrong here?

Re-examine for a moment the individual preferences of the senators. Assume that a certain amount of reasonableness and consistency exists in each senator's personal ordering of the desirability of the three proposals. In particular, assume that each senator's ordering is *transitive*. Transitivity means that if x, y, and z are any three alternatives and a senator prefers x over y and prefers y over z, then the senator must prefer x over z.

If an individual's preferences are transitive, then his preference list can be denoted in a convenient way. Suppose a senator finds proposal C most attractive, proposal B least attractive, and proposal A intermediate to the other two. We may denote the preference list by $(C\,A\,B)$. Then transitivity implies that one proposal is favored over another exactly if it appears to the left of the other in the list.

We now make our first demand on the decision-making process: the collective preference must also be transitive. We want to guarantee that whenever the group prefers x to y and prefers y to z, then it must also prefer x to z. It is on this imminently reasonable and apparently innocent demand that simple majority voting stumbles badly. At least since the time of the Marquis de Condorcet (1743−1794), those concerned about just voting procedures and mechanisms noted the possibility that intransitive social preferences could result when the variation of simple majority voting we described is applied to a list of individual transitive preferences.

To be specific, there are six possible ways an individual can rank-order the three proposals: $(A\,B\,C)$, $(A\,C\,B)$, $(B\,C\,A)$, $(B\,A\,C)$, $(C\,A\,B)$, and $(C\,B\,A)$. Suppose that the preferences of the senators on the dependents' allowance proposals break down as follows:

$$(A\,B\,C)\text{: 31 votes,}\quad (B\,C\,A)\text{: 34 votes,}\quad (C\,A\,B)\text{: 35 votes.}$$

To simplify this example, we assume that none of the other three orderings are represented.

Which proposal will be adopted? If the originally outlined procedure is followed, the Senate will first choose between A and B. Since 66 senators prefer A over B, B will be eliminated from consideration. A second vote will be taken between A and C. Now 69 senators will opt for C, and only 31 for A. Thus, the Senate would adopt Proposal C.

A loud objection can be expected from the advocates of proposal A. It has already been established, they would argue, that the Senate prefers A to B. It is also clear that in a direct vote between B and C, B would receive 65 votes so that the Senate certainly prefers B to C. But if the Senate prefers A to B and B to C, then it must prefer A over C to maintain transitivity.

As you have just seen, the Senate's normal procedures do not necessarily lead to transitive group preferences. But is transitivity always so important? In any legislative situation, it might be argued, the body always has at any moment only the option between two proposals. Only after one of the original two proposals is voted down in favor of the other may a third proposal be introduced.

If group transitivity is not guaranteed, however, more serious problems arise. The result of the legislative deliberation may depend, not on the individual wishes of the members or the inherent worth of the proposals, but on the *order* in which the proposals are offered for consideration. To illustrate with this same example, suppose the agenda is arranged so that A and C are the two original proposals discussed. When a vote is taken, C triumphs over A. When proposal B is finally introduced, it competes against C and wins, 65 to 35.

The author of each of the three proposals *A*, *B*, and *C* then has a legitimate argument that his or her proposal is the one that is "most favored" by the Senate as a whole. Although legislative bodies almost universally employ the modification of simple majority voting we've discussed here, we see that it fails to be a just one. The procedure yields nontransitive group preferences. It does not always produce the same collective preference given the same set of individual preferences. It is subject to manipulations by those who control the ordering of items on the agenda.

A suggested modification of simple majority voting when there are more than two alternatives is to conduct all possible two alternative elections, decide each one by majority vote, and declare as the winner the single alternative that beats all others. Such an alternative, which triumphs over every other choice in head-to-head balloting, is called the *Condorcet Winner*. While this is an appealing decision rule, it doesn't always work, as it is possible that no Condorcet Winner exists. In our Senate example, there is no Condorcet Winner, since *C* beats *A*, *A* beats *B*, and *B* beats *C*.

What procedure should be used if the group wishes to guarantee transitivity and to guarantee that the group decision is purely a function of the individual preferences? How is the "will" of the group to be determined? One possibility often suggested is to adopt a *weighted voting scheme*.

B. Weighted Voting

Weighted voting mechanisms are often used to score athletic, artistic, and beauty contests. The individual ratings of a collection of judges are pooled to determine the final overall rankings of the contestants. Preassigned numerical weights are attached to each first-place rating, each second-place rating, and so on. A contestant receives a score that is the sum of the weights of the opinions of the individual judges. The group ranking of the contestants is then determined by their total scores. The person with the highest number of points is the winner, the individual with the next highest number is the first runner-up, and so on down the list. Notice that this procedure can be employed in a situation either like Example 2 or like Example 3.

Example 4

An instructor offers a class of 25 students the option of a take-home final examination, an in-class final examination, or a major term paper. The class will choose the single option that every student will experience.

Table 6.1 shows how many students rated each option first choice, second choice, and third choice.

Table 6.1

	Take-Home Exam	In-Class Exam	Term Paper
First Choice	7	10	8
Second Choice	12	3	10
Third Choice	6	12	7

Suppose 4 points are given for each first-place vote, 2 for second place, and 1 for third. Then the points for each option and total points are displayed in Table 6.2. The term paper emerges as the winner.

Table 6.2

	Take-Home Exam	In-Class Exam	Term Paper
Points			
4 for First	28	40	32
2 for Second	24	6	20
1 for Third	6	12	7
TOTAL	58	58	59

But suppose a different number of points are designated for each place. If, for example, we give 10 points for each first-place vote, 7 for second, and 3 for third, then a different outcome arises. Table 6.3 shows the results. Here the take-home exam is the overall top choice.

Table 6.3

	Take-Home Exam	In-Class Exam	Term Paper
Points			
10 for First	70	100	80
7 for Second	84	21	70
3 for Third	18	36	21
TOTAL	172	157	171

This example demonstrates the first problem with weighted voting: the outcome may depend on exactly how many points are given for each place.

Is there an allocation of points under which the in-class exam wins? Take 10 points for first, 3 for second, and 1 for third. Table 6.4 shows the results.

Table 6.4

	Take-Home Exam	In-Class Exam	Term Paper
Points			
10 for First	70	100	80
3 for Second	36	9	30
1 for Third	6	12	7
TOTAL	112	**121**	117

There are even more serious weaknesses with a weighted voting scheme: *individual voters may have incentives to falsify their true preferences*.

Consider a beauty contest example in which there are four contestants, labeled w, x, y, and z and three judges. Judges 1 and 2 each rank the contestants in the order $(x\,y\,z\,w)$, while judge 3 ranks them $(z\,w\,x\,y)$. If 5 points are assigned for a first-placed ranking by a judge, 4 points for second, 2 for third, and 1 for fourth, then x earns 12 points, y and z each earn 9, and w earns 6. The winner is contestant x, while y and z tie for second, and w is last.

Suppose that between the time the judges' ratings are submitted and the winner is announced, it is discovered that y has broken the rules of the contest. He is disqualified. The scoring system is now applied to the remaining contestants. It should yield the same results, we believe, especially since y is inferior to x, according to the tastes of each individual judge.

Yet if y is deleted, the rankings become $(x\,z\,w)$ for judges 1 and 2 and $(z\,w\,x)$ for judge 3. Now when the weights are tabulated, x still has 12 points, but z has 13 and w has 8. The master of ceremonies dutifully declares z the winner of the contest. Needless to say, x is furious and his attorney sues the contest committee, claiming her client has been treated unjustly.

This weighted voting scoring mechanism violates an ethical value and poses a practical political problem. Whether a group believes x is better than z or not should be a judgment independent of the group's feelings about a third contestant y. Weighted voting does not preserve this independence.

Here is the practical political problem: In the example, judges 1 and 2 have given their true preferences in their ratings. They think x is best and would like to see x emerge as the eventual winner. Suppose that they have heard rumors that y has not been completely rigorous in following the rules. If judges 1 and 2 were to switch their ratings to $(x\,y\,w\,z)$, they would make it more likely for x to win over z in the event that y is disqualified. These two judges would be falsifying their own preferences.

A fair and equitable voting mechanism should not encourage such falsification. Each voter should feel secure in casting a personal ballot that lists the alternatives exactly in the order in which she would like to see the outcome. Weighted voting schemes remove this security.

The procedure of weighing places in individual preference orderings with numbers and using these numbers to find the societal ordering of proposals or candidates is attributed to Jean-Charles de Borda (1733–1799). His "Mémoire sur les Élections au Scrutin," published in 1781, was the first mathematical theory of elections. When confronted with the possibility that voters might mask their true preferences in order to help their favorite candidate emerge on top, Borda is reported to have replied "My scheme is only intended for honest men."

DEFINITION We call a voting mechanism *manipulable* if there is at least one voter who by disguising his true preferences may ensure a group preference ranking he prefers to the one that would have been obtained had he submitted his true preference ordering.

We have just seen that weighted voting is a manipulable mechanism. The terms *nonmanipulable*, *strategy-proof*, or *sincere* are used to describe voting mechanisms that are not manipulable.

Various other schemes have been proposed for determining a group-preference ranking from lists of individual preferences, and some of them are widely used. These include plurality voting, instant runoff voting, approval voting, range voting, and proportional representation. Each seems to suffer from one or another defect. The injustices of

these voting mechanisms raises the question of whether it is possible to design one that everyone will agree is just and democratic. If it is possible, what would the rules of such a voting procedure look like?

III. An Axiomatic Approach

What is a *just* voting mechanism? To answer this question, we begin by listing some conditions or axioms that a voting system might reasonably be required to satisfy if it is to be labeled a "fair" one. Once the set of axioms is set, we can ask mathematical questions. Is the set of axioms consistent? If so, how many different structures satisfy them? If the axioms are inconsistent, which ones should be eliminated or modified?

In the first place, the mechanism will be translating a list of individual voter preferences into a group-preference list. The voters may differ greatly in their likes and dislikes of candidates or proposals. We do not wish to restrict the freedom of any voter to state her true preferences. Accordingly, the first axiom looks like this:

AXIOM 1 (INDIVIDUAL SOVEREIGNTY) Each voter may order the candidates (or alternative proposals) in any way he or she chooses and may even indicate indifference between pairs of candidates.

The second axiom demands that the system always produce a societal judgment that is transitive and depends only on the individual ballots cast by the voters.

AXIOM 2 (EXISTENCE OF SOCIAL WELFARE FUNCTION) For every collection of lists of individual preferences, the mechanism produces a unique list of society's preferences. The society's preferences are transitive.

Note that Axiom 2 removes the inequities associated with simple majority voting when more than two alternatives are being considered. It also rules out some schemes that do guarantee transitivity. For example, one mechanism might be to put all the individual lists into a hat and draw out at random one of these, which will be designated society's preference list. Since the societal choice corresponds to some particular individual's, it will be transitive. This scheme would not satisfy Axiom 2, because a second implementation of the mechanism (drawing again from the hat) might result in a different outcome. The *uniqueness* feature of the societal list would be violated.

A decision procedure that simply makes the societal outcome the alphabetical listing of the alternatives or one that selects the eldest voter and declares that person's preferences to be the group's preference would be consistent with Axiom 2.

The third axiom is a weak constraint that has generated no controversy among voting-theory experts. It simply asks that in those cases in which everyone prefers x to y, so does the society.

AXIOM 3 (UNANIMITY) If every individual prefers one alternative to another, so does the society.

It would certainly be unreasonable to claim that the society's ranking reflected that of its members if no one agreed with it. In the context of a "Social Choice Function" where all we require is the determination of a winner (society's top choice), the principle of

Unanimity is usually called *Pareto efficiency*. A social choice function is Pareto-efficient if it chooses alternative a whenever a is at the top of every voter's list.

The weighted voting schemes discussed in Section II satisfy Axioms 1, 2, and 3; the proof is left to the reader. The fourth axiom is designed to eliminate the difficulties associated with such systems.

AXIOM 4 (INDEPENDENCE OF IRRELEVANT ALTERNATIVES) The social ordering of any pair of alternatives depends only on the preferences of the individuals between the members of that pair.

This axiom implies that if we want to know whether the society prefers x to y or y to x, we need only examine the relative rankings of x and y on each voter's preference list; we need not look at the rankings of any other candidates.

To understand this axiom better, return for a moment to the beauty contest example. If the rankings as originally turned in by the judges give a group judgment of x higher than z, then any other set of ballots in which judges 1 and 2 rank x higher than z and in which judge 3 ranks z higher than x will result in a group judgment of x higher than z if Axiom 4 holds. In other words, any two elections in which all voters preserve their preference between two particular candidates will yield the same group preference between *those* two candidates.

Let's illustrate this point with an example. The American Film Institute (AFI) selects a panel of three famous critics to choose the best movie ever produced. The critics are Roger, Janet, and Zoey and the three finalist films are *Citizen Kane*, *The Godfather*, and *Casablanca*. Each critic will submit a ranking of the three movies. A social welfare function will then be used to obtain an overall ranking of the three. The exact details of the particular social welfare function are unknown, but the AFI guarantees that it does satisfy Axiom 4.

We'll consider the relative rankings of *Citizen Kan*e and *Casablanca*. Suppose that the individual rankings of the critics are those shown in Ballot 1, and in the overall ranking, *Citizen Kane* winds up above *Casablanca*. If Axiom 4 holds, then *Citizen Kane* would also end up rated higher than *Casablanca* if the individual rankings are those of Ballot 2. This would be true because Roger and Janet rated *Citizen Kane* above *Casablanca* and Zoey listed *Casablanca* higher than *Citizen Kane* in both ballots. No individual voter changed a relative ranking of these two alternatives. Note that we do not know that the social ranking actually put *Citizen Kane* in a higher position than *Casablanca*; that is not material to the question of whether Axiom 4 holds.

Ballot 1

	Roger	Janet	Zoey
(1)	Citizen Kane	The Godfather	Casablanca
(2)	Casablanca	Citizen Kane	The Godfather
(3)	The Godfather	Casablanca	Citizen Kane

Ballot 2

	Roger	Janet	Zoey
(1)	Citizen Kane	Citizen Kane	Casablanca
(2)	The Godfather	Casablanca	The Godfather
(3)	Casablanca	The Godfather	Citizen Kane

A voting mechanism that satisfied Axiom 4 would ensure that voters gain nothing by disguising their true preferences.

It is very easy to design a voting mechanism that satisfies the first four axioms. Simply designate some particular voter as a dictator and decree that society's preference list will just be a copy of that one person's list. The reader should verify that this dictatorial mechanism is consistent with Axioms 1−4. Although having a dictator is certainly an extremely efficient voting mechanism, it is not what most people would call a "democratic" institution. The final axiom rules out such systems.

AXIOM 5 (NONDICTATORSHIP) There is no voter with the power that for all choices x and y, if he ranks x over y, then so does the society regardless of how other voters feel about x **and** y.

A dictator is a voter whose submitted preference list always becomes the society's preference list.

To make the axiomatic model realistic, assume that there are a finite number of individual preference lists. To make it interesting, assume that there are at least three different alternatives to be ranked. (The reader is asked to show that simple majority voting satisfies Axioms 1−5 if there are exactly two candidates or proposals being considered.)

These five axioms describe conditions all of which seem natural and desirable to demand of a voting mechanism. You may, in fact, believe that the axioms demand too little for the mechanism to deserve the adjective "democratic." The axioms do not demand, for example, that each voter's preference list be treated equally; some individuals might be given more "votes" than others. The axioms do not require that society prefer x to y if a simple majority prefers x to y. The axioms also do not insist that the same procedures be used on all pairs of alternatives. Conceivably, a mechanism that used a dictator to decide between Proctor and Swenton while using simple majority voting on Emerson vs. Peterson might be allowed.

The surprising fact is that even this "reasonable" set of axioms is inconsistent. The five demands are incompatible with each other. It is impossible to devise any voting mechanism that will simultaneously satisfy all of them.

This result is known as the General Impossibility Theorem. It was first stated by Kenneth J. Arrow in 1951 in a pioneering essay that sought to place voting theory on an axiomatic basis. Arrow's original proof contained a technical error and a correct proof was first supplied by Julian Blau in 1957. Arrow's Theorem has provoked a considerable amount of discussion by social scientists, philosophers, political theorists, and economists.

IV. Arrow's Impossibility Theorem

We state the theorem in a manner that is both provocative and that indicates the direction of its proof:

THEOREM (ARROW'S GENERAL IMPOSSIBILITY THEOREM) Axioms 1−4 imply the existence of a dictator.

The remainder of this section presents a proof of the theorem. Assume, then, that there is a voting mechanism satisfying Axioms 1−4. We need one additional definition.

DEFINITION A set V of individual voters is *decisive for alternative x against alternative y* if x is socially chosen by the voting mechanism whenever every individual in V prefers x to y and every individual not in V prefers y to x.

This concept is somewhat subtle and requires some explanatory remarks:

a. If the mechanism is a dictatorial one, then the dictator is a one-person set who is decisive for every pair of alternatives.

b. Axiom 3 on Unanimity asserts that the set of all voters is decisive for every pair of alternatives. Should every voter prefer x to y, then so would society. Of course, not every voter might share this preference between x and y.

If some of the voters prefer x to y and others prefer y to x, we need to know more about the details of the voting mechanism to determine the societal ranking.

c. Decisiveness is really a *potential* power. If V is a set that is decisive for x against y, then one of the conditions that must be present in order to predict that society prefers x to y is that everyone in the set V prefers x to y. If a particular individual belongs to V and he prefers y to x, then the fact that V is decisive for x against y does not really give V much influence on the outcome.

d. The other condition that must be met if decisiveness is to be used to predict a societal ranking is that all the individuals not in V must prefer y to x. If V is decisive for x against y, if everyone in V prefers x to y, and if someone not in V also prefers x to y, then we can make no accurate prediction about the societal ranking of x against y unless we have more detailed knowledge about the voting mechanism.

To clarify this point, suppose we have a society with seven members: Mike, Judy, Eli, Sherry, Abby, John, Sasha, and Anne. The voting mechanism is simple, but rather peculiar. The societal ranking is always exactly the opposite of Mike's preference ranking. Let x and y be any two alternatives. Then the set whose members are Eli, Abby, and Sasha is decisive for x against y. If these three prefer x to y and the other four members prefer y to x, then, in particular, Mike prefers y to x. Since Mike prefers y to x, the society prefers x to y. Consider, however, what happens if Eli, Abby, Sasha, and Mike all prefer x to y. Then society will prefer y to x, even though Mike has voted the same way as all the members of a decisive set.

e. A set V may be decisive for x against y but not necessarily decisive for y against x or decisive for any other pair of alternatives. This is due to the fact that Axioms 1−4 do not require the voting mechanism to operate the same way for all pairs of alternatives.

With these warnings about the notion of a decisive set in mind, we proceed to the main part of the proof of the theorem. The proof proceeds by verifying two claims:

Claim I There is some pair of alternatives and some individual who is decisive for that pair.

Claim II If an individual is decisive for some pair of alternatives, then he or she is decisive for every pair of alternatives—that is, the individual is a dictator.

Proof of Claim I For any pair of alternatives x and y, there is at least one nonempty decisive set—namely, the set of all individuals. Among all sets of individuals that are decisive for some pair of alternatives, pick a minimal set. This is a set V of voters and a pair of alternatives x, y so that V is decisive for x against y and no proper subset of V is decisive for *any* pair of alternatives. Axiom III on unanimity means that no empty set can be decisive of any pair of alternatives; thus, V contains at least one voter.

If such a minimal decisive set contains exactly one voter, then we are done with Claim 1. Hence, assume that V contains at least two voters. Let V^* be the set consisting of exactly one voter from V, $V^{\#}$ the subset of V consisting of all voters in V not in V^*, and let V' be the set of all voters in the society not in V. Now V^* is a proper subset of V. We shall show that V^* is decisive for some pair of alternatives, thus contradicting the minimality of V.

Suppose that V is decisive for x against y, and let z be any other alternative. Suppose that the voter in V^* ranks the alternatives $(x\,y\,z)$, all the voters in $V^{\#}$ rank them $(z\,x\,y)$, and all the voters in V' rank them $(y\,z\,x)$.

Note first that all voters in $V = V^* \cup V^{\#}$ prefer x to y and that all voters not in V prefer y to x. Since V is decisive for x against y, society prefers x to y.

Next note that $V^{\#}$ is smaller in size than V, so it is not decisive for any pair. In particular, $V^{\#}$ is not decisive for z against y. This implies that society prefers y to z, for otherwise we would have society preferring z to y when everyone in $V^{\#}$ does and no one outside $V^{\#}$ does.

Finally, use the transitivity of the societal preference. Society prefers x to y and y to z. Thus, society prefers x to z.

We then have one election in which V^* prefers x to z, everyone outside V^* prefers z to x, and the society prefers x to z. By Axiom 4 on Independence of Irrelevant Alternatives, society will prefer x to z *whenever* all individuals maintain these preferences between x and z. Hence, V^* is decisive for x against z. This contradicts the assumption that V is a minimal decisive set. The conclusion, then, is that minimal decisive sets contain precisely one voter. Claim I is verified.

Proof of Claim II Let J be some individual member of the society and write:

1. "$a\overline{D}\,b$" to mean that a is socially preferred to b whenever J prefers a to b regardless of the orderings of other individuals
2. "$aD\,b$" to mean that a is socially preferred to b if J prefers a to b and all other voters prefer b to a

These notations are useful since the condition of dictatorship is that $a\overline{D}\,b$ *for all pairs of alternatives* a and b, while aDb is true if and only if J is a decisive set for a against b.

To complete the proof of Claim II, the following lemma is useful.

LEMMA Suppose there are three alternatives a, b, c. Then

1. aDb implies $a\overline{D}c$, and

2. aDb implies $c\overline{D}b$.

Proof of Lemma Let J rank the alternatives $(a\,b\,c)$ and suppose everyone else ranks b higher than a and c. Since aDb, we conclude that society prefers a to b. Since all individuals prefer b to c, so does society. By transitivity, society prefers a to c. The axiom on Independence of Irrelevant Alternatives asserts that whenever J prefers a to c, so does society, regardless of how the other voters rank c and a. In other terms, $a\overline{D}c$.

To prove that aDb implies $c\overline{D}b$, suppose first that J ranks the alternatives in the order $(c\,a\,b)$ and all other voters rank them $(c\,b\,a)$ or $(b\,c\,a)$. Since aDb, society prefers a to b. By unanimity, society prefers c to a. Transitivity then gives a society preference of c over b. Applying Axiom 4 again, we have $c\overline{D}b$.

This completes the proof of the lemma.

We can now finish the proof of Claim II. Suppose xDy for some pair of alternatives x and y.

Case 1 There are exactly three alternatives: x, y, z.

We must show that $a\overline{D}b$ for all pairs of alternatives—that is,

$$\begin{array}{ll} (1)\ x\overline{D}z & (2)\ z\overline{D}y \\ (3)\ x\overline{D}y & (4)\ y\overline{D}z \\ (5)\ z\overline{D}x & (6)\ y\overline{D}x \end{array}$$

The proof of (1) follows directly from the lemma with $a = x$, $b = y$, and $c = z$. Similarly, (2) follows from a direct application of the lemma. Now that we know that $x\overline{D}z$, we also have xDz. Now apply the lemma with $a = x$, $b = z$, and $c = y$. The conclusions are that $x\overline{D}y$ and $y\overline{D}z$, giving (3) and (4). The proofs of (5) and (6) are left to the reader.

Case 2 There are more than three alternatives.

Suppose xDy holds and let a and b be any alternatives.

(i) If x and y are the same as a and b, add a third alternative z to x and y and apply the result of Case 1 to show that xDy implies $x\overline{D}y$ and $y\overline{D}x$. Hence, both $a\overline{D}b$ and $b\overline{D}a$ hold.

(ii) If exactly one of a and b is distinct from x and y, add it to x and y to form a triple and apply Case 1.

(iii) If both a and b are distinct from x and y, two steps are needed: First, add a to x and y; obtain $x\overline{D}a$ so that xDa. Second, consider the triple x, a, b; obtain $a\overline{D}b$.

Thus, xDy for *some* x and y implies $a\overline{D}b$ for *all* alternatives a and b. This completes the proof of Claim II and hence the proof of the theorem. ◇

Since Axioms 1−5 are inconsistent as they stand, any attempt to strengthen them—such as demanding that all voters be treated equally—will not remove inconsistency. A voting system that satisfies some of the axioms must violate some of the others. We will not enter here the heated argument as to which is the "best" axiom to modify or discard. The interested reader may follow the debate by consulting the References.

V. The Liberal Paradox and the Theorem of the Gloomy Alternatives

In the years since Kenneth Arrow formulated his Impossibility Theorem, social scientists, political theorists, economists, philosophers, and mathematicians have examined many aspects of the approach he pioneered. They have looked, for example, at how the axioms might be weakened or how they are interrelated to each other or how other fairness criteria might be formulated. In this section, we state and prove several theorems they have discovered that indicate that even more minimal requirements of fairness on a voting mechanism ensure that it is a dictatorial one.

A. The Liberal Paradox

One of the tenets of classical liberalism, extolled for example in John Stuart Mill's famous 1869 essay *On Liberty* is that each individual should be "locally decisive" with respect to a narrowly defined sphere that is that person's private concern. Society should not be able to decide, for example, which religion you must practice or which books you cannot read.

In the context of the problems Kenneth Arrow addressed, Amartya Sen formulated an *Axiom of Minimal Liberalism*: there are at least two individuals each of whom is strongly decisive for some pair of alternatives. A stronger *Axiom of Liberalism* would require that at least one such pair of alternatives exist for every individual, but Sen does not need such an assumption.

Sen discovered and proved the *Liberal Paradox Theorem*: There is no social decision function that satisfies Citizen Sovereignty, Unanimity, and Minimal Liberalism.

We outline a proof. Suppose we have a social decision rule that satisfies Minimal Liberalism. Then there are individuals, call them John and Amy, and alternatives x, y, z, and w such that John is strongly decisive for x and y while Amy is strongly decisive for z and w. We assume, for simplicity, that all four alternatives are distinct.

If our mechanism also satisfies Citizen Sovereignty, then it must deal successfully with a profile that contains these rankings by John and Amy:

John	Amy
. . .	. . .
w	y
. . .	. . .
x	z
. . .	. . .
y	w
. . .	. . .
z	x
. . .	. . .

Suppose, in addition, that every other individual also ranked w above x and y above z.

Since John is strongly decisive for the pair x and y and John ranked x above y, the mechanism must rank x above y.

If Unanimity is also assumed, then the mechanism ranks w over x and y over z. Hence, our result is that society ranks w above x, x above y, and y above z. Transitivity would require that society rank w above z. But Amy is strongly decisive for the pair w and z, and she ranked z above w; hence, the mechanism must also rank z above w. Thus, Citizen Sovereignty, Unanimity, and Minimal Liberalism force the mechanism to fail the requirement of transitivity that a social choice function must have.

B. Theorem of the Gloomy Alternatives

Recall first Arrow's fourth axiom on Independence of Irrelevant Alternatives (IIA). Suppose we have two different sets L and L^* of rankings by the individual voters where every voter ranks a above b in both L and L^*. The IIA axiom requires that whenever a finishes higher than b under L, then a must finish higher than b under L^*.

We are going to consider a weaker version of IIA that only deals with the case in which alternative a winds up at the top of the social ranking:

DEFINITION Suppose a and b are two alternatives and we have two different sets L and L^* of rankings by the individual voters where every voter ranks a above b in both L and L^*. A social choice function is *monotonic* if whenever a finishes first under L it also finishes first under L^*.

We also want to consider another desirable outcome for a voting mechanism: it should never produce a social preference ranking P such that there is another possible ranking P' that every voter prefers to P. This criterion is a variation of the Unanimity Axiom. A voting mechanism that satisfies this condition is called *Pareto-efficient* or *Pareto-optimal*. Pareto optimality is an important concept with many applications in game theory, engineering, and the social sciences. The term is named after Vilfredo Pareto, an Italian economist (1848−1923) who used the concept in his studies of economic efficiency and income distribution.

In 1977, Eitan Muller and Mark Satterthwaite demonstrated that in situations with more than two alternatives or candidates, insisting on both monotonicity and Pareto efficiency requires accepting a dictatorial mechanism. Let's state their result more explicitly and examine the proof:

The Muller-Satterthwaite Theorem If there are at least three alternatives and a social choice function is both Pareto-efficient and monotonic, then the social choice function is dictatorial.

Proof The following proof is an adaptation of Philip Reny's [2000] argument. Suppose there are N voters numbered $1, 2, \ldots, i, \ldots, N$. Let the preference ranking for voter i be labeled as P_i. Then the collection $P = \{P_1, P_2, \ldots, P_N\}$ is called a *profile*; we'll also use the term *set of ballots*.

There are five steps in the proof of the Muller-Satterthwaite Theorem.

Step 1. Let a and b represent any pair of distinct alternatives (candidates), and let f be a social choice function—that is, given any profile P of rankings, $f(P)$ is a single element in the set of alternatives. You may think of $f(P)$ as the winner of the election as determined by a social choice function that is monotonic and Pareto-efficient.

Consider a profile P where every single voter placed a at the top of the list and b at the bottom. Since f is Pareto-efficient, we would have $f(P) = a$.

Start with voter 1, and think about what would happen if we begin to change that voter's ranking *P1* by raising b one position at a time. Since f is monotonic, the social choice remains a as long as b is below a in Voter 1's ranking.

Eventually, we would move b to the top of Voter 1's list, where a would now be in second place. Who could be declared the winner with this new profile? If c is any third alternative, then a was above c for every individual's ranking in both the original and the new profile. By the monotonicity property, f can't choose c as the winner. Thus, either b is declared the winner or a remains the chosen alternative.

If a is still the winner, we repeat the same process with voters 2, 3, 4, and so on. We must eventually reach some voter k so that when b rises above a in Voter k's ranking, the social choice function names b the winner. If, to the contrary, a remains the winner after we have gone through every single voter, then we would have a profile in which b is ranked in top position by every voter and a sits in the second position. By Pareto efficiency, a could not be the winner.

Tables 6.5 and 6.6 show the situations immediately before immediately after we raise b above a in Voter k's ranking.

Table 6.5 Before switching *a* and *b* for Voter *k*.

P_1	. . .	P_{k-1}	P_k	P_{k+1}	. . .	P_N	Social Choice
b	. . .	*b*	*a*	*a*	. . .	*a*	
a		*a*	*b*				*a*
				b	. . .	*b*	

Table 6.6 After switching *a* and *b* for Voter *k*.

P_1	. . .	P_{k-1}	P_k	P_{k+1}	. . .	P_N	Social Choice
b	. . .	*b*	*b*	*a*	. . .	*a*	
a		*a*	*a*				*b*
				b	. . .	*b*	

Step 2. Now let's move alternative a to the bottom of the lists of voters $1, 2, \ldots, k-1$ and to the second from the bottom position for voters $k+1, \ldots, N$. Examine the parallel pictures in Tables 6.7 and 6.8.

Table 6.7 After switching a and b for Voter k.

P_1	. . .	P_{k-1}	P_k	P_{k+1}	. . .	P_N	Social Choice
b	. . .	b	a	.	. . .	.	
		.	b				a
				a		a	
a		a		b	. . .	b	

Table 6.8 After switching a and b for Voter k.

P_1	. . .	P_{k-1}	P_k	P_{k+1}	. . .	P_N	Social Choice
b	. . .	b	b	.	. . .	.	
.		.	a				b
				a		a	
a		a		b	. . .	b	

Let's determine the social choice under the profiles represented by Tables 6.7 and 6.8. Start with Table 6.8 and compare it with Table 6.6. No individual's ranking of b versus any other alternative has changed: b is first for voters $1, 2, \ldots, k$ and last for voters $k+1, \ldots, N$. Since f is monotonic and b was the winner for the profile of Table 6.6, b must also be the winner for the profile of Table 6.8.

Now compare Table 6.8 with Table 6.7. The social choice for Table 6.8 is b and f is monotonic, so the social choice in Table 6.7 could only be b or a. If the social choice for Table 6.7 were b, then monotonicity would also imply that the social choice for Table 6.5 would also have to be b. Since we've already shown that the social choice for Table 6.5 is a, we would have a contradiction. Thus, the social choice for Table 6.7 must be a.

Step 3. Consider now the profile of rankings shown in Table 6.9 where c is an alternative distinct from a and b.

We can obtain the profie of Table 6.9 from the Table 6.7 profile without altering the ranking of a versus any other alternative in any individual's ranking. By monotonicity, the social choice in Table 6.9 must also be a.

Step 4. Now examine the profile of rankings in Table 6.10 that we derived from Table 6.9 by interchanging the ranking of alternatives a and b for voters $k+1, \ldots, N$. This is the only difference between the two profiles.

The social choice for Table 6.9 was a and f is monotonic. Therefore, the social choice for Table 6.10 can only be a or b. We claim that the social choice

Table 6.9

P_1	. . .	P_{k-1}	P_k	P_{k+1}	. . .	P_N	Social Choice
.	. . .	.	*a*	.	. . .	.	
.		.	*c*	.		.	
.	.		*b*	.		.	
.	.						*a*
c	. . .	*c*	.	*c*	. . .	*c*	
b	. . .	*b*	.	*a*	. . .	*a*	
a	. . .	*a*	.	*b*	. . .	*b*	

Table 6.10

P_1	. . .	P_{k-1}	P_k	P_{k+1}	. . .	P_N	Social Choice
.	. . .	.	*a*	.	. . .	.	
.		.	*c*	.		.	
.	.		*b*	.		.	
.	.						*a*
c	. . .	*c*	.	*c*	. . .	*c*	
b	. . .	*b*	.	*b*	. . .	*b*	
a	. . .	*a*	.	*a*	. . .	*a*	

for Table 6.10 must also be *a*. Suppose, to the contrary, that it was *b*. Since alternative *c* is ranked above *b* in every individual's Table 6.10 ranking, monotonicity would imply that the social choice would remain *b* even if *c* were raised to the top of every voter's list, contradicting Pareto efficiency.

Step 5. Observe that alternative *a* is at the very top of Voter *k*'s list and the very bottom of every other voter's ranking. That means we can built an *arbitrary* profile of rankings with *a* again at the top of Voter *k*'s preferences without lowering the ranking of *a* versus any other alternative in any individual's ranking. Monotonicity would now imply that the social choice must be *a* whenever Voter *k* puts *a* at the top of his or her list. Thus, Voter *k* must be a dictator for alternative *a*.

Alternative *a*, however, was chosen arbitrarily, so we can conclude that for every alternative a^* there is some Voter k^* who is dictatorial for a^*. Clearly there cannot be distinct dictators for distinct alternatives, for if $k \neq k^*$, what is the social choice if Voter *k* lists alternative *a* first and Voter k^* lists alternative a^* first? Since each is a dictator, the social choice must be *a* and a^*. But there is only one social choice, so $a = a^*$ and thus $k = k^*$. There is a single dictator for all alternatives.

This completes the proof of the Muller-Satterthwaite Theorem. ◇

We now turn to the question of whether there exist perhaps some more basic principles of fairness in a voting mechanism that might themselves imply both monotonicity and Pareto efficiency.

In our examination of Weighted Voting, we saw this method was subject to *manipulation*: an individual voter might, by falsifying his true preferences, help obtain a social choice he preferred rather than the one that would have resulted had he submitted a ranking that truly reflected his preferences.

Social choice theorists have developed the idea of *strategy-proof* choice mechanisms to characterize systems not subject to such manipulation. We need one bit of notation to state the definition of strategy-proof clearly and concisely.

Suppose that $P = \{P_1, \ldots, P_i, \ldots, P_N\}$ is a profile of rankings and that P_i' is a ranking of Voter i different from P_i. Then (P_i', P_{-i}) denotes the profile obtained by replacing P_i in P with P_i', and $f(P_i', P_{-i})$ is the social choice under this new profile.

The idea of *strategy-proof* is that if a voter submits a ranking P_i' different from the one P_i he prefers and that false submission changes the social choice, then he likes the new choice less than the original choice. Here is the formal definition:

DEFINITION A social choice function f is *strategy-proof* if for every individual voter i and every possible profile P and every possible ranking P_i', $f(P_i', P_{-i}) \neq f(P)$ implies that individual i ranks $f(P)$ above $f(P_i', P_{-i})$ under P_i.

Recall also that the Unanimity axiom implies that each possible candidate could win the election (that is, finish at the top of the social ranking) if all voters ranked that candidate at the top of their individual preference lists. A condition apparently weaker than Unanimity is that for each candidate, there must be some set of individual preferences under which that candidate wins. This condition is called the *onto* property.

DEFINITION A social choice function is *onto* if for each alternative a there is at least one profile under which the social choice is a.

Our next theorem demonstrates that any onto, strategy-proof voting mechanism will automatically be monotonic and Pareto-optimal.

Reny's Theorem If a social choice function is strategy-proof and onto, then it is Pareto-efficient and monotonic.

Proof We follow closely here the proof by Philip Reny [2001]. We break the argument down to three steps.

Step 1. Suppose that the social choice for some profile P is alternative a, and that for every alternative b, the ordering P_i' ranks a above b whenever P_i does for some fixed voter i.

We want to show that $f(P_i', P_{-i})$ is also a. We will proceed using proof by contradiction.

Assume that $f(P_i', P_{-i}) = b$ for alternative $b \neq a$. Since f is strategy-proof, it must be true that P_i ranks a above b. On the other hand, alternative a's ranking does not fall in replacing P_i with P_i'. Hence, a, which was the choice under P,

must also be ranked above $b = f(P_i', P_{-i})$ in P_i'. But this result contradicts strategy-proofness. Thus, $f(P_i', P_{-i}) = f(P) = a$.

Step 2. In this step, we'll demonstrate that the social choice procedure is monotonic. Suppose now that the social choice for some profile P is alternative a, and that for every alternative b, the ordering P_j' ranks a above b whenever P_i does for *every* voter i. Now we can move from $P = (P_1, P_2, \ldots, P_N)$ to $P' = (P_1', P_2', \ldots, P_N')$ by changing the ranking of each voter i from P_i to P_i' one at a time, and because we have shown that the social choice must remain unchanged for every such change, we must have $f(P') = f(P)$. Thus, f is monotonic.

Step 3. In the final step, we'll show that the social choice procedure is Pareto-efficient. Let a be any alternative. Since f is onto, there is some profile P such that $f(P) = a$. Since f is monotonic, the social choice remains alternative a when a is raised to the top of every individual's ranking. But f being monotonic also implies that the social choice remains a regardless of how the alternatives below a are ranked by each individual. Consequently, the social choice is a whenever every individual ranks a at the top—that is, f is Pareto-efficient. Hence, Reny's Theorem is true.

We complete this section with what I call the *Theorem of the Gloomy Alternatives*: if we demand even the simplest conditions on a voting mechanism, then we must settle for a dictator or a manipulable system. Allen Gibbard and Mark Satterthwaite independently discovered and proved this theorem in the mid-1970s.

The Gibbard-Satterthwaite Theorem If there are at least three alternatives and the social choice function is strategy-proof and onto, then the social choice function is dictatorial.

Proof With the given hypotheses, the Reny Theorem tells us that the social choice function is Pareto-efficient and monotonic. The Muller-Satterthwaite Theorem then implies that it is dictatorial. ◇

VI. Instant Runoff Voting

In the light of Arrow's Impossibility Theorem and the results of Gibbard and Satterthwaite, how should we proceed to derive a group decision out of individual preference rankings when we're faced with more than two alternatives?

In this section, we will investigate one alternative, *Instant Runoff Voting*, or IRV for short, that is gaining in popularity. Instant Runoff Voting is a variation of what we might call *Classic Runoff Voting*. Classic Runoff Voting deals with a two-stage process wherein all voters at each stage indicate a single candidate as their top choice. If none of the candidates achieves a majority of the votes, there is a runoff election between the two top candidates.

The 2008 U.S. Senate race in Georgia provides a recent example. In the general November election, the incumbent Republican Senator Saxby Chambliss was the leading candidate with 49.8% of the votes. His Democratic opponent Jim Martin garnered 46.8%.

Allen Buckley of the Libertarian party won 3.4%; two other write-in candidates received a handful votes. The total vote was 3,752,577. A month later, a runoff election between Chambliss and Martin took place. Chambliss won with 57.4% of the 2,137,956 votes cast. Note that voter turnout for the runoff contest in December was substantially smaller than the original numbers for the November election.

In Section I, we saw that runoffs between the top two alternatives may result in situations in which a majority of the voters actually prefer the eliminated third candidate to the winner of the runoff. This sort of result will occur because modified simple majority voting does not always guarantee transitive results. In addition to the theoretical shortcomings, classical runoff voting also poses a number of practical difficulties. Runoff elections are expensive for a city or state to conduct. At the local level a municipal election may cost several hundred thousand dollars; for a state, the outlay may be in the millions of dollars. The candidates must also raise additional money to campaign for an additional month or 6 weeks after the first balloting. Voters must wait that additional period before knowing who will represent them as the winner. Finally, runoff elections typically attract far fewer voters than the first run. In the Georgia example, almost half of those who voted in the initial contest failed to cast a ballot in the runoff.

Instant runoff voting solves many of these problems. Under IRV, each voter submits an individual preference list ranking all the candidates rather than simply indicating a first choice. Then first-place choices are tabulated. If a candidate receives a majority of first choices, that candidate is elected. If no candidate receives a majority of first choices, the candidate receiving the fewest first choices is eliminated. Ballots cast for the eliminated candidate are now counted toward those voters' second choices.

This process continues until one candidate receives a majority and is elected. Once all the preference lists have been submitted, a computer can easily carry out the successive rounds of transferring of votes until a majority winner is found. The election result can be determined shortly after the polls close, with no need of additional campaigning or for additional trips by voters to the ballot box. Thus, Instant Runoff Voting does solve the practical problems that plague classic runoff election processes.

To see in more detail how Instant Runoff Voting works, consider the following example. Suppose we have six different groups of voters of different sizes and four candidates Marc, Rhonda, Julie, and Brian. There are a total of 1,150 voters; a candidate must tally more than 575 votes to win. Table 6.11 displays the voter groups, their sizes, and their shared preferences.

Thus, Marc has 400 first-place votes, Rhonda has 300, Julie has 250, and Brian has 200. Under plurality voting, Marc would win the election. Under classic runoff voting, there would be a second-round election between Marc and Rhonda. In that contest, the 450

Table 6.11 Initial Preference Rankings for IRV Example

Group	I	II	III	IV	V	VI
Size	400	150	150	250	150	50
1st	Marc	Rhonda	Rhonda	Julie	Brian	Brian
2nd	Rhonda	Julie	Marc	Brian	Julie	Julie
3rd	Julie	Brian	Julie	Rhonda	Rhonda	Marc
4th	Brian	Marc	Brian	Marc	Marc	Rhonda

voters in Groups I and VI would vote for Marc while the 700 voters in Groups II, III, IV, and V would opt for Rhonda, making Rhonda the winner.

Instant runoff voting produces a different result. There is no majority winner in Round One, so candidate Brian is eliminated, as he got the lowest number of first-place voters. The 200 people who listed Brian at the top of their preference lists now have their votes transferred to their second-place candidate. Table 6.12 displays the redistributed ballots at the start of Round Two. In this case, Julie gets all 200 votes.

Now Marc still has 400 votes and Rhonda retains her 300, but Julie now has 450. Julie has the lead, but still lacks of majority. We thus end Round Two with no winner. Rhonda has the fewest votes and we eliminate her, moving her votes with the next highest candidate as we enter Round Three. Julie will get 150 votes from Group II, and Marc gets an equal number of votes from Group III. Table 6.13 shows the redistributed ballots for Round Three.

At this point, candidate Julie has 600 votes and candidate Marc has 550. Julie now has a majority and is declared the winner.

In addition to solving the practical difficulties classic runoff elections pose, Instant Runoff Voting has other appealing features. Some voting theorists argue that IRV satisfactorily address the issue of the "third-party spoiler" or "wasted vote" scenario. This is the problem that frequently occurs in many states that don't have runoffs, but that declare as victor the plurality winner. Here is a typical situation: suppose there are three candidates for governor: a Republican, a Democrat, and a Progressive who is to the left of the Democrat on the political spectrum. None of these three is the favorite of a majority of the voters. Each voter can cast only one ballot for a single candidate.

To be specific, suppose the electorate falls into three categories as shown in Table 6.14:

Table 6.12 Redistributed Ballots for Round Two

Group	I	II	III	IV	V	VI
Size	400	150	150	250	150	50
	Marc	Rhonda	Rhonda	Julie	Julie	Julie
	Rhonda	Julie	Marc	Rhonda	Rhonda	Marc
	Julie	Marc	Julie	Marc	Marc	Rhonda

Table 6.13 Redistributed Ballots for Round Three

Group	I	II	III	IV	V	VI
Size	400	150	150	250	150	50
	Marc	Julie	Marc	Julie	Julie	Julie
	Julie	Marc	Julie	Marc	Marc	Marc

Table 6.14

Group	I	II	III
Size	44%	36%	20%
	Republican	Democrat	Progressive
	Democrat	Progressive	Democrat
	Progressive	Republican	Republican

Some of the Group III voters simply prefer the Progressive candidate to the Democrat, while others are also seeking to build the Progressive Party by demonstrating it represents a significant part of the electorate. Early public opinion polls show 44% support for the Republican, 36% for the Democrat, and 20% for the Progressive. In response to the polls, many Group III members fear they would be "wasting" their votes by putting the Progressive at the top of their lists. Their anxiety is that doing so would give the victory to the Republican, their last choice. Similarly, pressure begins to build on the Progressive candidate to drop out of the race. The Democrats say the Progressive is a "spoiler": he can't win and by staying in the race prevents the Democrat, who is the first or second choice of every voter, from winning. History shows that in such situations, most of the Group III voters will cast their ballots for the Democrat in the hope perhaps of electing "the lesser of two evils." The Progressive winds up with only one or two percent of the vote. Political pundits then conclude that there is only very minimal support for a Progressive agenda when, in fact, about one in five people favor it.

Adherents of the Instant Runoff Voting system contend that under their system, the Group III voters can submit their true preferences without fear of hurting their second choice's chances of ultimately winning if their first choice is eliminated. If their third party favorite, the Progressive, does tally 20% of the first-place votes in the initial round, then the party will be taken more seriously by the public, Progressives will be included in future candidate debates, the media will pay more attention to them, and the party might attract more supporters. In a similar fashion, any small third party might benefit from IRV.

In a short 1871 paper "Application of Mr. Hare's System of Voting to the Nomination of Overseers of Harvard College" in the *Journal of Social Science*, the American architect William Robert Ware first introduced Instant Runoff Voting in the form we have described. He was building on an idea of the Australian Thomas Hare, who proposed a related method, called the *single transferable vote*, for electing multiple members to a governing board in a manner that reflected proportional representation among many different constituencies.

Since its first use in Australia at the turn of the 20th century, Instant Runoff Voting has spread to a number of other countries. IRV is used to elect members of the Australian House of Representatives, the President of Ireland, the national parliament of Papua New Guinea, and the Fijian House of Representatives. The Labour and Liberal Democrat parties in the United Kingdom use IRV to select their leaders.

In recent years, IRV has gained much attention in the United States and has been adopted for various elections in many states, including Arkansas, California, Colorado, Florida, Illinois, Louisiana, Maryland, Massachusetts, Michigan, Minnesota, New Mexico, North Carolina, South Carolina, Vermont, and Washington. Instant runoff voting is sometimes called *alternative voting* in the United Kingdom, *preferential voting* in Canada and Australia, and *ranked choice voting* in the United States.

Californians for Electoral Reform summarizes the major arguments in favor of IRV, claiming that it

a. Results in majority rule

b. Eliminates the "spoiler" dilemma, wherein voting for a weak favorite candidate causes one's least favorite candidate to win

c. Allows for a diverse candidate field while also ensuring that the winner has the support of a majority coalition

d. Encourages positive campaigns, because candidates depend on the second choices of voters for other candidates

e. Works cheaply and conveniently, because it collects all the information necessary to determine a majority winner on one ballot

The California group also argues that by ranking candidates, "voters are able to express their true preferences without worrying about wasting their votes or spoiling the election and helping elect their least favorite candidate. For this reason alone, IRV often leads to higher turnout and stronger democracy. Candidates need to build a base of first choice support, but also reach out to the broader voting population in order to be acceptable to the majority."

With so many arguments in favor of IRV, are there good arguments against it? Arrow's Impossibility Theorem implies that IRV cannot satisfy Axioms 1−5. As an illustration of what can go wrong with IRV, let's examine an example with 21 votes whose preferences among four candidates split into four groups whose sizes and numbers are shown in Table 6.15.

The winner needs a majority of 11 or more votes to win. At the conclusion of Round 1, we that Karzi has 7, Barak has 6, Chavez has 5, and Patel has 3. Karzi has the plurality, but falls short of a majority, so Patel is eliminated, and the first choice votes of Group IV members are transferred to Chavez to start Round 2. Table 6.16 shows the resulting situation: Karzi has 7 votes, Barak has 6, and Chavez has 8.

There's still no candidate with a majority. The IRV rules require that we eliminate Barak, redistribute the votes of Groups III and IV, to their next choice and go on to Round 3. Table 6.17 displays the result.

At this point, Karzi has 13 and Chavez has 8. Karzi commands the majority and is declared the winner.

Note that Karzi won, despite the fact that all the voters in Group IV originally ranked Karzi at the very bottom of their lists. To see what difficulties are associated with IRV, examine what should happen if these voters had a change of heart just before the election and moved candidate Karzi to the very top of their preference lists. We expect that Karzi

Table 6.15

Group	I	II	III	IV
Size	7	6	5	3
1st	Karzi	Barak	Chavez	Patel
2nd	Barak	Karzi	Barak	Chavez
3rd	Chavez	Chavez	Karzi	Barak
4th	Patel	Patel	Patel	Karzi

Table 6.16

Group	I	II	III	IV
Size	7	6	5	3
	Karzi	Barak	Chavez	Chavez
	Barak	Karzi	Barak	Barak
	Chavez	Chavez	Karzi	Karzi

Table 6.17

Group	I	II	III	IV
Size	7	6	5	3
	Karzi	Karzi	Chavez	Chavez
	Chavez	Chavez	Karzi	Karzi

Table 6.18

Group	I	II	III	IV
Size	7	6	5	3
1st	Karzi	Barak	Chavez	Karzi
2nd	Barak	Karzi	Barak	Patel
3rd	Chavez	Chavez	Karzi	Chavez
4th	Patel	Patel	Patel	Barak

Table 6.19

Group	I	II	III	IV
Size	7	6	5	3
1st	Karzi	Barak	Barak	Karzi
2nd	Barak	Karzi	Karzi	Barak

should still win, since Karzi is at least as highly rated on everyone's ballot as she was originally. Table 6.18 shows the revised preference lists.

In Round 1, Karzi now has 10 votes, which is not quite a majority, so IRV dictates that we must eliminate Patel, who didn't get any first-place votes, and then Chavez.

After transferring the appropriate votes, we see (Table 6.19) that Karzi retains the 10 votes, but candidate Barak now has 11, which is a majority. Barak wins. This outcome seems perverse: how could Karzi lose an election if more people rank her first? This problem with IRV is often described as the More-Is-Less Paradox: If the winner were ranked higher by some voters, all else unchanged, then another candidate might have won.

Note also that with the original preference rankings, a majority of voters prefer Barak to Karzi, a majority prefer Barak to Chavez, and a majority prefer Barak to Patel, yet Karzi won under IRV. Thus, IRV can lead to what sometimes called the Thwarted Majorities Paradox: a candidate who can defeat every other candidate in a direct-comparison majority vote may not win the election!

We'll illustrate one more problem of IRV, illustrated by an example that comes from Peter Fishburn and Steven Brams. Imagine a municipal election with three candidates: Bitt, Huff, and Wogg. Two of the voters, Mr. and Mrs. Smith, are on the way to the polls when their car breaks down; they are thus unable to register their preferences. Both of them favored Bitt to Huff to Wogg and would have turned in that ranking. "Although they liked Mrs. Bitt best," write Fishburn and Brams, "they were almost as fond of Mr. Huff, but disliked and mistrusted Dr. Wogg."

When the votes were tabulated the next day, it was discovered that 1,608 people had turned in preference lists. Table 6.20 shows the number received for each of the six possible orderings.

Table 6.20

Group	I	II	III	IV	V	VI
Size	417	82	143	357	285	324
1st	Bitt	Bitt	Huff	Huff	Wogg	Wogg
2nd	Huff	Wogg	Bitt	Wogg	Bitt	Huff
3rd	Wogg	Huff	Wogg	Bitt	Huff	Bitt

A candidate needs 805 votes to win. The preference lists submitted show 499 votes for Bitt, 500 votes for Huff, and 609 for Wogg. IRV then requires that we eliminate Bitt and transfer her votes to the other candidates. The 417 votes of Group I go to Huff, making his total 917, and Group II's 82 votes go to Wogg, increasing his total to 691. Huff is the winner. When the Smiths read the result in their newspaper, "they were delighted that Dr. Wogg had not won. They did feel a twinge of regret that their friend, Mrs. Bitt, was beaten. Perhaps their votes would have made a difference."

They certainly would have made a difference, but not the way the Smiths hope. Had they made it to the polls in time, Mrs. Bitt would have had 501 votes at the end of the first round. IRV would have eliminated Mr. Huff, who only had 500. When Huff's votes are transferred, Group III's 143 go to Bitt, giving her a new total of 644, but all of Group IV's 357 votes would go to Wogg, raising his total to 966, well above that required for a majority. The Smiths' well-intentioned votes to help Bitt or Huff would have backfired and made Wogg the winner!

Fishburn and Brams dub this particular problem of IRV the No-Show Paradox: The addition of identical ballots with a particular candidate ranked last may change the winner from some other person to that particular candidate.

VII. Approval Voting

While Instant Runoff Voting has its strong adherents and a number of cities and states have adopted this mechanism, it also has strong critics who are eager to point out some of its potential pitfalls. Peter Fishburn and Steven Brams created the examples we have just seen, associated with the fictional town of Bramburn, that demonstrate some of these problems.

What method of social choice do Fishburn and Brams advocate? They propose an entirely different option for voters. In traditional voting, each individual can only indicate her *single top choice*. For the Borda count, Instant Runoff Voting, or the more general schemes envisioned by Arrow, each voter submits *a ranked ordering* of all candidates. Fishburn and Brams suggest giving each voter the option of voting for any number of

Used with permission of FairVote.org

FIGURE 6.1 The logos for two advocacy groups.

Table 6.21

Group	I	II	III	IV	V	VI
Size	17	40	40	20	46	4
1st	Ford	Ford	Olds	Olds	Saab	Saab
2nd	Olds	Saab	Saab	Ford	Ford	Olds
3rd	Saab	Olds	Ford	Saab	Olds	Ford

candidates for a given office. The candidate who collects the most votes wins. This system is called Approval Voting. In plurality voting, the direction "Vote for one candidate" would appear above the list of contenders. In Approval Voting, the direction would be "Vote for as many candidates as you like."

To illustrate how Approval Voting works and how it may result in outcomes different from Instant Runoff Voting or Plurality Voting, consider the rankings displayed in Table 6.21 for an election among three candidates (Ford, Olds, and Saab) where there are 167 votes, falling into six different groups.

Under traditional plurality voting, Olds wins 60 first-place votes, followed by Ford with 57 and Saab with 50. With Instant Runoff Voting, we would eliminate Saab and transfer 46 votes to Ford and 4 votes to Olds. After the transfer, Ford has 103 votes and Olds has 64, making Ford the winner.

To determine the outcome under Approval Voting, we need to know some additional information. Suppose that all voters check off the names of their top two candidates. Then Ford wins 57 votes from Groups I and II, while picking up an additional 66 votes from Groups IV and V; Ford's total is 123. Now Olds has $17 + 40 + 20 + 4 = 81$ votes. In this case, Saab gets 40 votes from Group II, 40 from Group III, 46 from Group V, and 4 from Group 4. Here 130 voters listed Saab, so Saab is the winner.

For a different scenario, suppose 25 members of Group II list both Ford and Saab, while 15 just list Ford, everyone in Group VI only lists Saab, and everyone in Group I lists both Ford and Olds. In addition, suppose that each of the remaining groups splits in two, half listing their top choice only and half listing their top two choices.

Table 6.22 then shows the number of votes each candidate receives from members of each of the groups. Saab remains the winner with 95 votes.

Another way to tabulate votes under approval voting is to list all the possible subsets of candidates and the number of ballots for each of these. For the example we have just been looking at, we have

Ford	15
Olds	$20 + 10 = 30$
Saab	$4 + 23 = 27$
Ford, Olds	$17 + 10 = 27$
Ford, Saab	$25 + 23 = 48$
Olds, Saab	20
Ford, Olds, Saab	0

Table 6.22

Group	I	II	III	IV	V	VI	Total
Ford	17	40	0	10	23	0	90
Olds	17	0	40	20	0	0	77
Saab	0	25	20	0	46	4	95

Table 6.23

Subset	Ballots	Subset	Ballots
None	1, 100	Olds, Saab	1425
Ford	10, 738	Olds, Dodge	1824
Olds	6561	Saab, Dodge	608
Saab	7626	Ford, Olds, Saab	148
Dodge	8521	Ford, Olds, Dodge	5605
Ford, Olds	3578	Ford, Saab, Dodge	143
Ford, Saab	659	Olds, Saab, Dodge	89
Ford, Dodge	6679	All	523

Source: From Steven J. Brams and Jack H. Nagel, "Approval Voting in Practice," *Public Choice* **71** (1991): 1–17.

From this table, we calculate the total for each candidate:

$$\text{Ford} : 15 + 27 + 48 = 90$$
$$\text{Olds} : 30 + 27 + 20 = 77$$
$$\text{Saab} : 27 + 48 + 20 = 95$$

Advocates for Approval Voting claim that it:

- Is simple
- Is easy to understand
- Is practical to implement
- Increases voter turnout
- Helps elect the strongest candidate
- Gives voters more flexibility
- Gives minority candidates their proper due

Table 6.23 shows another example with four candidates. The data represent actual votes cast under Approval Voting in the 1988 president election of the Institute of Electrical and Electronic Engineers (IEEE). The IEEE is an international organization with more than 200,000 members. In the 1988 poll, 55,310 members returned ballots. We've changed the names of the candidates to Ford, Olds, Saab, and Dodge.

Table 6.24

Group	Ford	Olds	Saab
80	X		X
15		X	X
5		X	X
	80	20	100

Table 6.25

Group	I	II	III
Size	80	15	5
1st	Ford	Olds	Saab
2nd	Saab	Saab	Olds
3rd	Olds	Ford	Ford

To determine a candidate's total, we need to add the votes for all subsets to which that candidate belongs. For example, to find Ford's total:

$$10{,}738(\text{Ford}) + 3{,}578(\text{Ford, Olds}) + 659(\text{Ford, Saab}) + 6{,}679(\text{Ford, Dodge})$$
$$+ 147(\text{Ford, Olds, Saab}) + 5{,}605(\text{Ford, Olds, Dodge}) + 143(\text{Ford, Saab, Dodge})$$
$$+ 523(\text{All}) = 28{,}073$$

Similar calculations for the remaining three candidates gives these totals:

Olds: 19,753

Saab: 11,221

Dodge: 23,992

One disadvantage of Approval Voting over Instant Runoff Voting is that voters have no way of indicating a strong preference for one candidate and a weaker one for another candidate if they are willing to accept either of them as an eventual winner. Real voters almost always will have different degrees of support for different candidates. Approval Voting forces individuals to cast equally weighted votes for candidates whose names they check off on the ballot.

Consider, for example, a set of 100 ballots submitted in an Approval Voting situation. Table 6.24 displays the results that the election officers would see at the end of the day. Eighty voters approve of Ford and Saab, 15 approve of Olds and Saab, and 5 approve of Olds and Saab. Thus, Ford gets a total of 80 votes, Olds gets 20, and Saab gets 100. Approval Voting makes Saab the winner.

Of the 80 voters who checked off the names Ford and Saab, we have no idea how many preferred Ford to Saab, how many had the opposite ranking, or how many might have been indifferent between the two. Suppose it were possible to see the preference rankings of our voters? Table 6.25 shows a possible set of rankings; each voter opted to check off the names of his top two choices.

We see from Table 6.25 that candidate Ford was the top choice of an overwhelming majority (80%) of the voters. Surely, any reasonable voting process should give the victory to candidate Ford. Note that candidate Ford wins under Plurality Voting, under Instant Runoff Voting, and wins in a sequence of two candidate Simple Majority contests. Approval Voting has a serious deficiency: a candidate who is the first choice of a staggering majority of voters might not win the election!

The Center for Voting and Democracy finds this flaw in Approval Voting to be a serious one. The Center also notes that Approval Voting does not solve the spoiler problem. Voting for your second choice candidate can in some cases lead to the defeat of your favorite candidate. In the example shown in Tables 6.24 and 6.25, if 55 of the 80 voters in Group I had checked only Candidate Ford and not displayed approval for Saab, then Ford would have won.

As a practical consequence, each candidate might benefit by encouraging her supporters to "bullet" vote—that is, only check her name as acceptable. If everyone did this and all voters complied, then Approval Voting reverts back to Plurality Voting.

Despite some of its theoretical shortcomings, Approval Voting is gaining acceptance as a new societal decision making method especially as an alternative to Plurality Voting. Approval Voting was used as early as the 13th century to select the doge (chief magistrate) of the Venetian Republic. It began to be studied seriously by voting theorists, economists, political scientists, and operations researchers in the middle 1970s when five scholars rediscovered it independently. Robert J. Weber of Northwestern University coined the term "Approval Voting." The selection of the Secretary General of the United Nations uses Approval Voting, and variations of it have occurred in elections in the Soviet Union, 19th-century England, and the American colonies in the 17th century. Some presidential straw polls and statewide referenda in the United States have also employed Approval Voting. Several major professional mathematical societies (Mathematical Association of America, American Mathematical Society, Institute for Operations Research and Management Sciences, American Statistical Association) use Approval Voting to choose officers, as do the Public Choice Society, the Society for Judgment and Decision Making, the International Joint Conference on Artificial Intelligence, the Econometric Society, and the National Academy of Sciences.

Instant Runoff Voting and Approval Voting are both seen as superior to Plurality Voting and appear to be the leading candidates to replace it. There is no agreement concerning which is superior. IRV supporters are quick to point out the shortcomings of Approval Voting, whose adherents are equally swift in noting that IRV fails some fairness criteria. You can follow some of the debate along with the successes and failures having one of these methods adopted by governments by checking their respective websites: www.FairVote.org and www.ApprovalVoting.org. The arguments can become quite heated at times: the mathematician Donald Saari, who has studied voting procedures extensively, once described Approval Voting as an "unmitigated disaster."

VIII. Topological Social Choice

A. Topological Social Choice

In previous chapters, we have often presented *discrete* and *continuous* models of the same situation. Our view of the social choice problem so far has been a discrete one: a finite

collection of voters examining a finite set of candidates. In this section, we will briefly discuss continuous analogs. In particular, we'll look at situations with a finite number of voters but a continuum of alternatives. The techniques and results in this field, *topological social choice*, were pioneered by the mathematician and economist Graciela Chilchilinksy. In this section, we will present a brief nontechnical discussion of topological social choice theory and some of its major findings so far. The field is in active development with new results unveiling connections previously hidden. Proofs of the principal theorems generally require tools and techniques of algebraic topology and thus lie beyond the mathematical prerequisites of this volume.

The fundamental problem of social choice theory is how to aggregate a collection of individual preferences among a set of alternatives into a single collective choice by a procedure that satisfies a number of prescribed fairness conditions. In our discussions so far, we have only examined situations involving a finite group of individuals (the *voters*) and a finite field of alternatives (the *candidates*). Our main result, Kenneth Arrow's Impossibility Theorem, deals with this discrete framework.

As an example, consider Dave and Judy's selection for a site to build their dream home. They have just purchased a large piece of property that contains a circular lake of radius 1 mile. The couple agree that they want their new house to be located on the lakeshore but disagree as to where the location should be. Four possible sites A, B, C, and D have been identified; they are shown in Fig. 6.2.

We want to identify some choice rule that will take Dave and Judy's preferences and select one of the four sites. For simplicity, let's assume that Dave and Judy are only asked to submit their top choice, the site each likes the best. What are some conditions we might want to impose on choice rules? We certainly want the rule to handle any combination of top choices submitted by our two voters and to output one of the four possible sites. We might also want to restrict our choice rule by insisting that it satisfy three additional constraints:

Unanimity: If all the voters choose the same site, then the choice rule also picks this site.

Anonymity: The choice rule treats all voters impartially. The same collection of profiles—whether they have the voters' names on them or not, whether we change the names of the voters on the ballots or not—always produces the same outcome. The choice rule pays attention only to the profiles submitted, not to which voter turned in which profile.

Stability: If one voter changes her opinion and now claims that her favorite site is next to her previous choice, then the output of the choice rule changes at most to a site next to the previous output—that is, choice rule selects the same site it did before or to one adjacent to it.

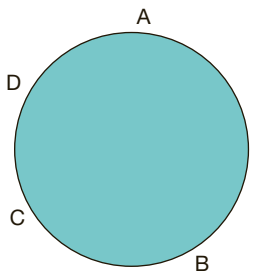


FIGURE 6.2

Table 6.26

		DAVE'S	TOP	CHOICE	
		A	**B**	**C**	**D**
JUDY'S	A	*A*	*A*	*B*	*A*
TOP	B	*A*	*B*	*C*	*D*
CHOICE	C	*B*	*C*	*C*	*C*
	D	*A*	*D*	*C*	*D*

Thus, stability implies that if the choice rule originally chose B when Judy submitted D and David submitted A, then the choice rule will choose A, B, or C if Judy submits A or C and David still submits A. The stability condition is meant to ensure that the social choice doesn't move much under small "errors" of the voters. Table 6.26 shows the assignments of choice rule that has stability.

Note that the ways we have formulated these conditions of unanimity, anonymity, and stability do not restrict the number of voters or the number of alternatives. They could equally well apply to a social rule that picks the best of n possible home sites taking into account the preferences of k family members.

It is an interesting exercise to show that in the case of Dave and Judy, it is not possible to construct a social choice rule that satisfies stability, anonymity, and unanimity if there are five sites available. In fact, Yuliy Baryshnikov (1993) proved a more general theorem that if $n > 2k$, there is no stable, anonymous, unanimous social choice rule.

To move to a *continuous* model of social choice, suppose Dave and Judy are free to choose any spot along the shore to locate their dream home. Each of their first-place choices is then a point on a unit circle; the social choice rule must look at their choices and assign some point on that circle. Set up a standard coordinate system with the origin at the center of the lake. Note that we can identify each point P on the circle with the directed line segment from the origin to P. This vector has length 1 and defines an angle θ between the positive-horizontal axis and the vector, measured in a counterclockwise fashion. We can thus describe the location of a point on the unit circle by giving its pair of Cartesian coordinates or simply by stating the angle θ. Dave and Judy each have infinitely many choices for θ. See Fig. 6.3.

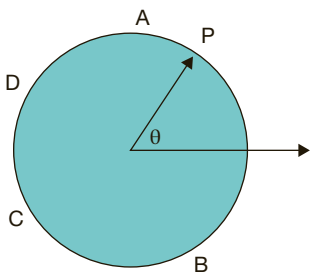


FIGURE 6.3

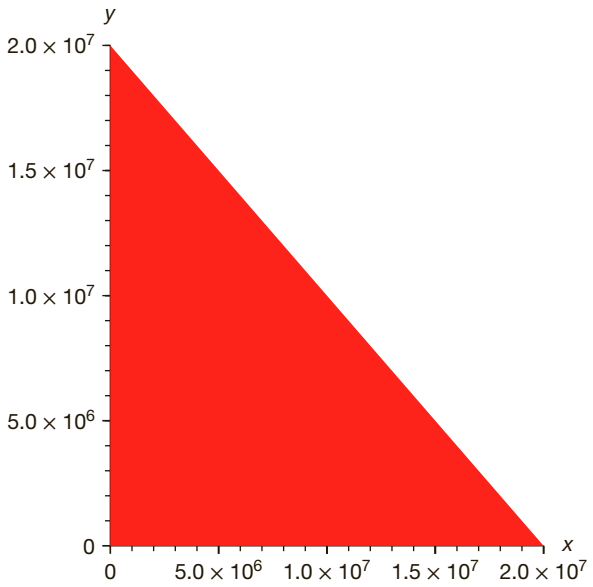


FIGURE 6.4

There are many situations in which the society needs to make choices among infinitely many possibilities. Consider for example a city council allocating a budget among different departments. For a very simple case, suppose the council has decided it can spend up to \$20 million and wishes to split the budget between education (E) and municipal services (M). The council may then choose any pair of nonnegative numbers E and M whose sum does not exceed \$20 million. The set of possible choices is equivalent to the set of points (x, y) in the first quadrant whose sum is less than or equal to 20,000,000—that is, $P = \{(x, y) : x \geq 0, y \geq 0, x + y \leq 20,000,000\}$. Fig. 6.4 illustrates this set.

In reality, the city council must distribute its budget over many departments. If there are n departments, then the set of all possible choices is equivalent to the set of points in Euclidean n-dimensional space all of whose coordinates are nonnegative and that sum to no more than 20,000,000—that is, $P = \{(x_1, x_2, x_3, \ldots, x_n)$: each $x_i \geq 0, x_1 + x_2 + \cdots + x_n \leq 20,000,000\}$.

In a series of papers beginning in the 1980s, Graciela Chichilnisky introduced a *topological* approach to social choice theory—that is, an approach anchored in the concept of *continuity*. Informally speaking, we assume first that if an individual prefers alternative x to alternative y, then that person prefers alternatives that are sufficiently close to x to alternatives close to y. We also want to consider social choice rules that follow a similar rule: if two profiles of preferences are sufficiently close together, then the social choices from the two profiles should also be close. Note that we will need some precise way to talk about "close" and "sufficiently close."

In beginning calculus, we study continuous functions between sets of real numbers. In more advanced classes, we investigate continuous functions between subsets of Euclidean spaces; we may, for example, examine a function that assigns a point in

Euclidean three-dimensional space E^3 to each point of the plane or perhaps a function that assigns an n-dimensional vector to each m-dimensional vector. A function f that assigns an element of a set B to every element of a set A is called a *continuous function from A to B* if, roughly speaking, every pair of sufficiently close points in A are sent to a pair of very close points in B. More exactly, suppose we have distance measures d_A and d_B for the sets A and B, respectively. The function f is *continuous at a point x* in A if for every $\varepsilon > 0$, there is a $\delta > 0$ such that $d_B(f(x), f(y)) < \varepsilon$ whenever $d_A(x, y) < \delta$. We say that f *is continuous on A* if it is continuous at each point of A.

A *neighborhood* of an element p is the set of elements whose distance from p is less than r for some positive number r, called the *radius* of the neighborhood. An equivalent definition of continuity is that for every neighborhood V of $f(p)$, there is a neighborhood U of p such that every element of U is sent by f to some element of V. The mathematical discipline *topology* studies continuity in Euclidean spaces of all dimensions and in more abstract spaces

We can view the idea of *continuity* in the infinite space of candidates as an extension of *stability* in the discrete case. With stability, a small change in a preference profile results in at most a relatively small choice in the output of the social choice rule.

In the Chichilnisky model, the possible alternatives (candidates) form a subset A of Euclidean n-dimensional space. She assumes each member of society has preferences over the set A, which vary smoothly as we move from one spot in A to another. She also assumed that preferences are *unsatiated*—that is, given any neighborhood U of a point x in A, there is some point y in U that you will prefer to x. We could then form a vector starting at x that points in the direction of greatest increase in our preference. For consistency purposes, we normalize the vector to have length 1.

Graciela Chichilnisky and Geoffrey Heal found a geometric description of when it is possible to solve problems like Dave and Judy's where we have k citizens instead of just two. They proved in 1983 that when each voter specifies an element from a space $\mathscr{B}$ of preferences, then a social choice rule that outputs an element of $\mathscr{B}$ for every k-tuple of elements of $\mathscr{B}$ can be constructed that is continuous, anonymous, and unanimous if and only if $\mathscr{B}$ is *contractible*. Although contractible is a technical term with a very precise meaning, you can think of it as meaning that there are no holes in $\mathscr{B}$ or that $\mathscr{B}$ does not surround a hole. Alternatively, a contractible space is one that can be continuously shrunk to a point inside itself.

In the Dave and Judy example with all points on the lakeshore under consideration, the social choice function assigns to each pair of angles $(\theta_{Judy}, \theta_{Dave})$ another angle, the output of the social choice rule. The set of all possible pairs of angles can be represented geometrically by a torus, the surface of a hollow doughnut. Since the preference space $\mathscr{B}$ in this case is a circle and the circle surrounds a hole, the circle is not contractible. Hence, there is no social choice rule in this situation that is unanimous, anonymous, and continuous.

Luc Lauwers (2000) calls the Chichilnisky-Heal result "The Resolution of the Social Choice Paradox" and observes that Chichilnisky's introduction of a topological approach to social theory "caused a major breakthrough in the disentanglement of the possibilities and limitations of preference aggregation. . . . Necessary and sufficient conditions to resolve the social choice paradox were established and new insights in the relationships between different aggregation axioms were obtained."

Are there reasonable situations in which the space of preferences would be contractible? One result suggests that if there is at least some "limited agreement" among the voters, then it is possible to have a continuous, unanimous, anonymous social choice rule. For example, if there is some fixed preference v on the circle no individual has, then the space of preferences is contractible.

Chichilnisky was successful in translating many of the important axioms of discrete social choice procedures into the continuous setting and discovering their geometric nature and how they related to certain classic results in topology. Her breakthroughs paved the way for many others to explore topological choice theory. Recently, Yuliy Baryshnikov (1993, 1997) discovered new, deeper connections between the discrete and continuous approaches to collective decision making. In his words, this work

> *demonstrates a remarkable interplay between two theories of social choice: a* topological *one, initiated and developed mainly by Chichilnisky, and the classical,* combinatorial *one, stemming from the work of Arrow. Both theories deal with the aggregation of preferences with apparently cardinally different notions of preferences. Recall that in the classical theory, the preferences are assumed to be given on discrete sets of alternatives and constitute a discrete set by themselves. This bounds the technique of the theory to be combinatorial. In the topological theory of social choice, the set of alternatives is assumed from the beginning to have the structure of a topological space . . .*
>
> *Until recently it has been implicitly assumed that the theories coexist but do not have much in common. The combinatorial setting was considered as primary and more natural and simple, while the topological one was usually seen as hi-tech fortresses with no life nearby. . . . Both theories are in fact much closer to each other than was commonly thought. Actually, I believe that they are in fact two different guises of the same theory which uniformly covers both discrete and continuous phenomena of the social choice theory.*

We don't yet have a complete unifying theory that satisfies everyone, but progress is being made. You can follow new developments in such journals as *Social Choice and Welfare*, *Theory and Decision*, *Economic Theory*, and *Voting Matters*. Perhaps you can make some discoveries yourself.

IX. Historical and Biographical Notes

A. Pliny the Younger

The question of what procedure to follow when a group of individuals must choose among more than two alternatives goes back in history at least as far as ancient Rome. In A.D. 105, Pliny the Younger recounts a decision facing the Roman Senate. The issue concerned the fate of the freedmen of the consul Afranius Dexter, who had recently died. The freedmen were former slaves whom Afranius had liberated and who were working as his paid servants. The senators considered three options: let these servants go free, banish them to a remote island, or execute them. (Roman practice was to execute the slaves immediately on the death of the master.)

Public domain

Pliny the Younger (61−113 A.D.), was a Roman lawyer, author, and natural philosopher. He witnessed and wrote about the eruption of Mount Vesuvius in August 79 during which his uncle and mentor Pliny the Elder died. His letters about Vesuvius were so keenly detailed that modern vulcanologists describe that type of eruption as Plinian.

Pliny was the presiding officer of the Senate. He favored leniency for the freedman, but he knew that those who supported his position, although they were the largest group numerically, did not command a majority. Thus, he called for a plurality vote, asking each of those who favored a particular outcome to go to a separate corner of the room. The head of the faction favoring execution quickly realized that the freedman would then be released to live as citizens of Rome. He persuaded his followers to drop their first choice and vote for banishment, which then commanded a majority.

B. Jean-Charles Borda

Jean-Charles, chevalier, de Borda (May 4, 1733−February 19, 1799), was a French mathematician, physicist, political scientist, and mariner. Born into a French aristocratic family, he spent much of his life as a naval officer and military engineer. At age 23, Borda wrote an important paper on projectile motion that led to his election as a member of the French Academy in 1764.

In 1770, he proposed the ranked preferential voting system now known as the *Borda count*. The French Academy used Borda's method for many years until Napoleon abolished it when he came to power in 1801.

Public domain

Jean-Charles de Borda

Borda fought with the French in the American Revolution, eventually commanding a fleet of six ships until the British captured him in 1782. Later in his career, he did considerable work on hydraulics and also helped define the meter as one ten-millionth of the distance from the North Pole to the Equator.

C. Marquis de Condorcet

Borda proposed his weighted voting as an alternative to the method advocated by the Marquis de Condorcet. Condorcet's most important work was an 1785 treatise *Essai Sur L'application de L'analyse À La Probabilité Des Décisions Rendues À La Pluralité Des Voix* (*Essay on the Application of Analysis to the Probability of Majority Decisions*). This 500-page work furthered the development of probability theory and laid out more completely a mathematical basis for social choice procedures. Condorcet proposed that the winning candidate should be the one who beats all other candidates in head-to-head elections. Such a candidate, as we have noted before, is called a Condorcet Winner. If, for example, there are four candidates A, B, C, and D and A defeats B in a simple majority runoff between the two, and similarly A defeats C and A defeats D, then A should be the winner. Borda agreed that if there was a Condorcet Winner, then it should be the group's choice, but thought it was impractical to insist on a Condorcet Winner, since there would be many ways that no candidate might be qualified.

Condorcet realized that there could exist situations where no such candidate exists and that intransitive results could result in such cases; we saw several examples in earlier sections of this chapter. He argued, however, that we should only consider social choice mechanisms that guarantee selecting the Condorcet Winner, if one exists. Because weighted voting does not guarantee the selection of such a candidate (see Exercise 66), Condorcet strongly opposed Borda's method.

Although Condorcet was responsible for seeing that Borda's paper, presented orally in 1770, was published at the same time his own *Essai* went to press, he could be quite contemptuous of his fellow Frenchman. Condorcet said Borda "likes nothing better than to waste his time drawing up prospectuses, examining machines, etc., and especially because, realizing he was eclipsed by other mathematicians, he abandoned mathematics for petty experiments. . . . Some of his papers display some talent, although nothing follows from them and nobody has ever spoken of them or ever will."

Condorcet's criticism of Borda was not an entirely an objective assessment. They had clashed earlier about what should be done with funds worth about $50,000 in today's dollars that the king owed the French Academy. Condorcet felt they should be used to pay his salary, whereas Borda felt they should be used to support experimental research.

Public domain

Marquis de Condorcet

Marie-Jean-Antoine-Nicolas de Caritat took his title Marquis de Condorcet from the town of Condorcet in Dauphiné, where his family resided. Born September 17, 1743, Condorcet was educated in Jesuit schools in Reims and then at the Collège de Navarre and the Collège Mazarin in Paris.

Condorcet displayed early talent in mathematics and was elected to the French Academy of Sciences at age 26. A 1772 volume on calculus he wrote was described by the eminent mathematician Lagrange as "filled with sublime and fruitful ideas."

Although he served in the administration of Louis XVI, Condorcet strongly supported the French Revolution. He advocated economic, religious toleration, abolition of slavery, free and equal public education, constitutionalism, and equal rights for women. After the Revolution, he served in the Legislative Assembly as a representative from Paris. Condorcet aligned himself with the moderate Girondists, who were ousted by the more radical Jacobins led by Robespierre.

Condorcet argued strongly against the new, hurriedly written, constitution that was drawn up by the Jacobins to replace the one that he himself had been chiefly responsible for drawing up. "Condorcet was no politician," wrote one of his biographers. "His uncompromising directness of manner and inability to suffer illogical windbags in silence made him many enemies and few friends." When a warrant for his arrest was issued, Condorcet went into hiding for half a year and eventually tried to flee from Paris but was caught and imprisoned on March 27, 1974. Two days later he was found dead in his prison cell. Whether he committed suicide, died from natural causes, or was murdered is still not known.

While he was in hiding, Condorcet wrote *Esquisse d'un tableau historique des progrès de l'esprit humain* (*Sketch for a Historical Picture of the Progress of the Human Mind*), now considered one of the major Enlightenment texts. Condorcet held a strong belief that human progress was linked to scientific discoveries and to mathematical and logical reasoning. He argued that there was an intimate connection between scientific advances and the spread of justice and human rights. His vision was that we could, through rational thought and the accumulation and sharing of knowledge, continually progress toward a utopian society.

D. Charles Dodgson (Lewis Carroll)

Most of the world knows him as Lewis Carroll, the author of the children's classic *Alice's Adventures in Wonderland*, its sequel *Through the Looking-Glass*, and the nonsense poems "The Hunting of the Snark" and "Jabberwocky." But Charles Lutwidge Dodgson (January 27, 1832–January 14, 1898) has an independent reputation as an Oxford University mathematician who made some serious contributions to social choice theory.

Public domain

Charles Lutwidge Dodgson (Lewis Carroll)

Table 6.27

Group	I	II	III	IV	V	VI	VIII
Size	2	2	2	2	2	1	1
1st choice	D	B	C	D	A	A	D
2nd choice	C	C	A	B	B	D	A
3rd choice	A	A	B	C	C	B	B
4th choice	B	D	D	A	D	C	C

In the mid-1870s, Dodgson wrote three short pamphlets about voting procedures to deal with multiple-candidate elections: *A Discussion of the Various Methods of Procedure in Conducting Elections* (1873), *Suggestions as to the Best Method of Taking Votes, Where More Than Two Issues Are to Be Voted On* (1874), and *A Method of Taking Votes on More Than Two Issues* (1876). Dodgson's practice as a mathematician was to develop his own solution to problems without considering previous work on the subject. Thus, it was unlikely that he had read the papers of Borda and Condorcet; he essentially came up with some of their approaches entirely on his own.

Dodgson suggested first looking for a candidate who was the first choice of a majority of voters. If such a candidate exists, then that person is declared the winner. If no one commands a majority, then Dodgson proposes examining all the two-candidate elections and seeing whether a Condorcet Winner emerges. If there is no Condorcet Winner, then Dodgson discusses various ways to proceed. These include versions of Instant Runoff Voting and the Borda count. He also proposed that the individual preference lists of the voters be examined to determine the smallest number of switches of consecutive candidates required to produce a Condorcet Winner. He envisioned a sequence of rounds of voting wherein the electors would be made aware of this information and offered the opportunity to submit new preference rankings.

Today, social choice theorists call an alternative a Dodgson Winner if it can be made a Condorcet Winner by interchanging as few adjacent alternatives in the individual rankings as possible. Consider, for example, the preference rankings of 12 voters among four candidates summarized in Table 6.27.

Candidate *D* has a plurality of five first-place votes, but no candidate commands a majority. Nor is there even a Condorcet Winner. If any three of the voters in Groups I–IV interchange their consecutive rankings of *C* and *A*, then *A* would become a Condorcet Winner. No other interchange by three or fewer voters of adjacent candidates in their rankings produces a Condorcet Winner. Hence, candidate *A* would be the Dodgson Winner.

E. Kenneth Joseph Arrow

The most prestigious and coveted international honors are the annual Nobel Memorial Prizes. These awards are given for outstanding achievements in medicine, literature, peace, chemistry, physics, and—since 1969—economics. The announcements of these awards each autumn are front-page news.

The Swedish Academy of Science selected Kenneth J. Arrow as a co-winner of the 1972 Nobel Memorial Prize in Economics. The academy cited Arrow's pioneering

contributions to general economic equilibrium theory and welfare theory. Although Arrow has made several key breakthroughs in economic theory, many of his colleagues rate the Impossibility Theorem of this chapter as his major achievement. According to the well-known economist Paul Samuelson, himself a Nobel Laureate in 1970, this theorem "is not only a stellar contribution to economics, it is as well a breakthrough for political science, and I would dare assert, for philosophy itself."

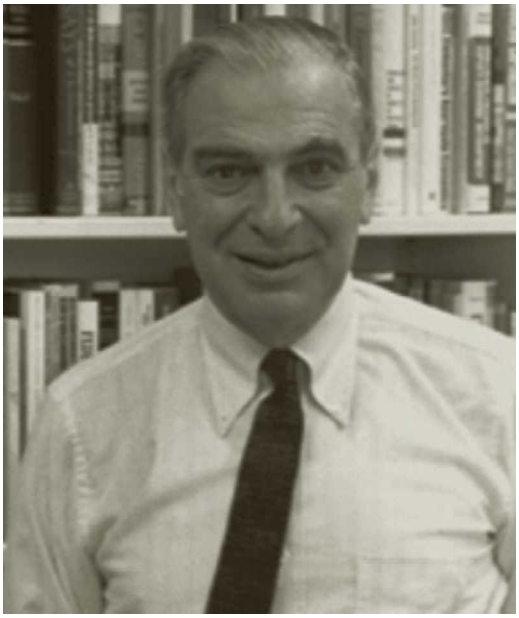


Photo courtesy of Kenneth Arrow

Kenneth J. Arrow

Arrow was born in New York City to Jewish immigrants raised on the Lower East Side. He graduated from the City College of New York in 1940 at the age of 18 with as he put it "a degree of Bachelor of Science in Social Science but a major in Mathematics, a paradoxical combination that was prognostic of my future interests." Arrow's advanced degrees were taken at another Manhattan institution, Columbia University. After a 4-year stint in the U.S. Army Air Force during World War II, Arrow was a research associate with the Cowles Commission at the University of Chicago from 1947 to 1949. The Impossibility Theorem was part of his Ph.D. thesis and in finished form was published as a book, *Social Choice and Individual Values*, in 1951.

Arrow began his work on the social welfare problem by trying to develop a reasonably fair function that took a collection of individual preference rankings and produced a group ranking. "I just started playing around," he told one interviewer. "It took me about two days to decide I was on the wrong track because I was looking for some solution. It didn't occur to me that there was no solution."

In 1949, Arrow joined the faculty of Stanford University where he taught for almost 20 years and was a major force in developing at Stanford an outstanding group of economic theorists and mathematical model builders. He also worked briefly with the Council of Economic Advisers during the administration of President John F. Kennedy. In 1968 Arrow moved to Harvard University where he became the James Bryant Conant University

Professor in 1974. In 1979, he returned to Stanford University with the position of Joan Kenney Professor of Economics and Professor of Operations Research. Arrow formally retired in 1991 but continues to be an active participant in economics conferences and was a vigorous bicyclist well into his mid-eighties.

Arrow has written or edited many books and dozens of papers whose topics include the mathematical theory of inventory and production, time series analysis of interindustry demands, linear and nonlinear programming, public investment and optimal fiscal policy, the theory of risk bearing, and general competitive analysis.

"Despite the deep abstraction of his econometric theories, friends consider Professor Arrow basically a humanist, a scholar who has always tried to apply fundamental theory to such social problems as medical care, education, race discrimination and water resources" wrote Robert Reinhold [1972] in a *New York Times* profile.

In appraising the work for which Arrow received a Nobel Prize, Samuelson [1972] wrote, "Men have always sought ideal democracy—the perfect voting system. . . . What Kenneth Arrow proved once and for all is that there cannot possibly be found such an ideal voting scheme. The search of the great minds of recorded history for the perfect democracy, it turns out, is the search for a chimera, for a logical self-contradiction. . . . Aristotle must be turning over in his grave. The theory of democracy can never be the same . . . since Arrow."

Despite his many honors and the demands of his research, Arrow has been remarkably available to undergraduate students and younger colleagues. He was the only senior faculty member at Harvard, for example, who volunteered to take on an assignment to lead discussion sections of an introductory economics course, a task usually delegated to graduate student teaching assistants. Four of Arrow's graduate students have themselves been awarded the Nobel Prize in Economics.

President George W. Bush presented the National Medal of Science, the nation's highest scientific honor, to Kenneth Arrow in November 2005 for "groundbreaking contributions to the pure theory of economics but [he] also holds a broad understanding of the social science arena in which theories are confronted and practical lessons worked out. His fundamental research on risk perception and behavior under uncertainty, and on equilibrium in markets with imperfect information, began a revolution in the design and analysis of market allocation mechanisms."

Photo by Eric Draper, Courtesy of the George W. Bush Presidential Library & Museum

President George W. Bush about to present the National Medal of Science to Kenneth Arrow

F. Amartya Sen

Amartya Kumar Sen, discoverer of the Liberal Paradox, is a Bengali economist and philosopher born on November 3, 1933. He won the Nobel Memorial Prize in Economic Sciences in 1998 "for his contributions to welfare economics" for his work on famine, human development theory, welfare economics, the underlying mechanisms of poverty, and political liberalism.

From 1998 to 2004, Sen was Master of Trinity College at Cambridge University, becoming the first Asian academic to head an Oxbridge college.

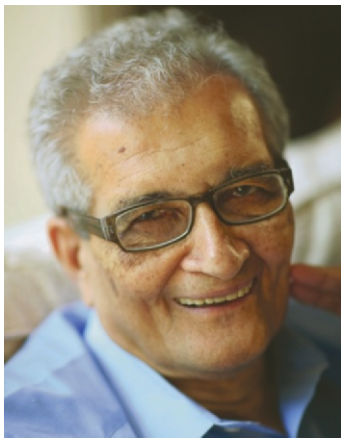


Photo by Graeme Robertson. Used with permission of SIPA USA

Amartya Sen

Among his many contributions to development economics, Sen has produced work on gender inequality. He is currently the Lamont University Professor at Harvard University. Sen's books have been translated into more than thirty languages.

G. Graciela Chichilnisky

One of the world's preeminent mathematical economists, Graciela Chichilnisky has achieved great success despite numerous obstacles. She is currently UNESCO professor of economics and mathematics and professor of statistics at Columbia University, where she also directs the Program on Information and Resources and the Center for Risk Management.

Chichilnisky was born March 27, 1946, in Buenos Aires, Argentina. Her parents Salomon Chichilnisky and Raquel Gavensky came from families fleeing Russian anti-Semitic pogroms at the turn of the 20th century. Her father battled his way from being a dock worker to becoming a medical doctor, then a professor of neurology, and later the national Secretary of Health, in which role he built many hospitals and a large part of Argentinean National Health system.

Chichilnisky excelled at school despite several anti-Semitic incidents. As a high school junior, she informally took courses at the University of Buenos Aires, where she gravitated to mathematics from an initial focus on philosophy and sociology.

> *"I wanted to do mathematics that would be applied to resolve social problems," she recalls. "I thought that studying Mathematics first would give me a control of the 'technology' that economists use to validate their theories and their policies. I felt it was important to 'control the*

technology'—rather than 'be controlled by it'—since many economists appear to fear the mathematical foundations of economics and adopt theories or policies based on what they learn from others mathematical models. I always liked the idea of creating my own mathematical models, rather than adopting somebody else's. Mathematics was a pleasure to learn. I think of it as the natural language that the brain uses to communicate with itself."

Just as Chichilnisky was about to start college, a military junta took over Argentina and closed down the universities. Fortunately she was invited to begin graduate level studies in mathematics at the Massachusetts Institute of Technology. She arrived in Cambridge, a single mother with infant child, without much knowledge of English, competing with doctoral students at a leading university without benefit of an undergraduate degree. Soon, however, she was at the top of her class with scholarship support from the Ford Foundation. She transferred to the University of California at Berkeley when she completed her Ph.D. in mathematics with her thesis "Group Actions on Spin Manifolds," a topic at the intersection of algebraic topology and physics.

Chichilnisky's first major job was as Director of Modeling at Fundacion Bariloche in Argentina, where she created a mathematical model of the world economy with an interdisciplinary team of prominent Latin American scientists, including geologists, sociologists, population experts, computer scientists, political scientists, and economists. In this model, she created the concept of "development based on the satisfaction of basic needs" rather than a more traditional approach of development through maximizing Gross Domestic Product. A radical idea at the time, Basic Needs was adopted as the cornerstone of efforts to define sustainable development by 166 nations at the Earth Summit in 1987.

Seeking a better understanding of international markets, Chichilnisky decided to return to Berkeley to work on a second doctoral degree, this time in economics. Her primary advisor was another Nobel Laureate, Gerard Debreu. While completing her dissertation, she also worked as a research associate at Harvard with Kenneth Arrow, who became an important mentor for her. After a period teaching mathematics and economics at Harvard while beginning her breakthrough research on topological choice theory, Chichilnisky moved to New York in 1977. There she served as director of research for the United Nations Institute for Training and Research while teaching at Columbia and Harvard.

Chichilnisky has published more than 16 books and 300 research papers covering a broad range of concerns, including oil in the international economy, development and global finance, information and uncertainty in markets, equity and efficiency in environmental markets, the gender gap, sustainability, dynamics, and uncertainty. She has received many honors and awards during her career. The Greek government named her Global Citizen of the Year in 2007, the University of Oslo conferred the Leif Johansen award on her in 1995, and *Hispanic Business* listed her as one of the most influential Latinos in the United States in 2006. She has frequently been mentioned as a possible future winner of the Economics Nobel Prize.

Chichilnisky has worked extensively on the Kyoto Protocol process, creating and designing the concept of the carbon market that has become international law in 2005. Working closely for several years with negotiators of the United Nations Framework Convention on Climate Change, the organization in charge of deciding world policy with respect to global warming, Professor Chichilnisky acted as a lead author of the

Intergovernmental Panel on Climate Change. The IPCC received the 2007 Nobel Prize for their work in this area.

Photo reproduced by permission of Graciela Chichilnisky

Graciela Chichilnisky

The road to such achievements has often been a very bumpy one for Chichilnisky. "Somewhere along this path," she writes, "I met uncontrollable forces, full of sound and fury, that thrust me up close into the stormy transition of women's roles at the turn of the 21st century." Her mentor at MIT suggested that she would not be able to compete successfully with students in the program there and suggested she transfer to a less demanding discipline at a weaker university. Later there were false malicious rumors that each of her doctoral theses were not original works, but had been written by her adviser or other senior colleagues. "As a foreign woman and a single mother, I started to face bewildering circumstances," Chichilnisky writes:

> *I learned a great deal during this process, about how women's intellectual property is treated in the masculine world of academia. . . . All this made me aware in the years that followed of the plights of other women in Ivy League universities whose intellectual work had been stolen or duplicated with impunity, or attributed to others. . . . While it is hard enough to compete with men in academic research, obtaining credit for what one has accomplished proved to be much more difficult . . . academic citations are consistently biased against women. Men resist*

giving women the basis for measuring academic achievement. They successfully deny them credit for their work in the form of academic citations. This problem is true today and in my case it has only become worse as my work has become better known.

Hostility escalated . . . as I grew professionally, it became relentless. Some of my colleagues recommended to my students not to work with me, others wrote threatening letters to my sources of funding and published numerous articles against my work; others acting as editors limited the ability of authors to write extending my work in this area. The campaign against my work extended to the Columbia administration, and created a hostile climate in which it was very difficult to work. . . .

At Columbia I have excellent colleagues who tried to stop this trend, but the forces of darkness succeeded. . . . Eventually my assistants' offices at the Economics Department were destroyed and my own office made unusable, my courses were removed, I was marginalized and treated with hostility, and my salary became so low that years later it had to be almost doubled and still remained below the average of male full professors.

When Chichilnisky discovered that her salary was substantially lower than her colleagues, she sought to remedy the discrepancy in what became "a David and Goliath epic." She describes the administration's response as one of "indifference and scorn." After failing to resolve the issue internally, Chichilnisky filed a lawsuit against Columbia in 1990. A settlement was reached in 1995 awarding her $500,000 and commitments from the university to provide the Program on Information and Resources, which Chichilnisky directed, with space and other support. By 2000, she came to believe that Columbia was not honoring its promises, was harassing her in retaliation for challenging the university, and was continuing to pay her unfairly. She filed suit again. Columbia filed counterclaims alleging she was delinquent in her duties and had secretly operated a private consulting company. After nearly 8 years of legal wrangling, the parties reached anew settlement in June 2008 awarding Chichilnisky $200,000. "The exact number isn't as important as the principle that it was a substantial amount of money that the university had to pay," she said. "And that has to do with who is right and who is wrong."

Despite the obstacles and hostility Chichilnisky has had to face, she remains positive:

For a woman to survive and to thrive she must learn to turn negative responses into positive resources. This is a perverse reversal to the Pavlovian response. I call this, for short, "turning dung into fertilizer." I believe it is one of the most important elements for women's success and happiness. It is a wonderful recipe for dealing with the "glass ceiling", a well-known and somewhat cruel situation where the more you succeed, the more you get punished. Think of it this way—energy is energy—and simply changing the sign of the response one receives from negative to positive allows one to use all the energy received constructively, and turn it into a survival tool. In mathematical terms, this is "life modulo two." It is the absolute value of the response that counts, not the sign . . . the only genuine source of happiness in life is the feeling of being useful to others. Nothing else does the job. This is true for anybody. It is not achievement or success, it is not money. It is this feeling of being useful that counts.

EXERCISES

I. Three Voting Situations

1. Show that a voting mechanism that gives a satisfactory resolution of situations in which a single best alternative must be chosen (as in Example 1) can be modified to handle situations when a full ranking of various alternatives is required (as in Example 2).

2. Are there any voting situations essentially different from those described in Examples 1−3? How are the outcomes determined in such situations?

II. Two Voting Mechanisms

3. What safeguards protect minority rights in systems using simple majority voting?

4. Suppose the senators are split in the following manner:
(A B C) 49 votes
(B C A) 49 votes
(C A B) 2 votes
If modified simple majority voting is used here, will the judgments of the Senate be transitive?

5. If the senators are split in the following manner:
(A B C) 32 votes
(B C A) 33 votes
(C A B) 35 votes,
show that there is a Condorcet Winner.

6. If 51 senators share the preference ranking (A B C), show that the Senate will have transitive preferences. Is there a Condorcet Winner?

7. Is it necessary for a majority of senators to share a common preference ranking to guarantee that the Senate judgments will be transitive? Why?

8. Consider a legislative body that only passes resolutions if they are supported by more than two-thirds of the members. How would such a body settle questions like those proposed by Examples 2 and 3? What inequities does such a system possess?

9. In order to correct past discrimination, it has been proposed that for a limited period, the votes of women be given twice the consideration of the votes of men—that is, each woman receives two votes on every proposal while each man receives one. Proposals are adopted or rejected on a simple majority count. Can intransitive results emerge? What other injustices are associated with such a system?

10. There are 1,000,000 shares of stock in the Emerson Construction Company. Two shares are owned by Mrs. Emerson, and the remaining shares are split evenly between her two sons. In deciding company policies, each shareholder has a number of votes equal to the number of shares he controls. How much relative power does Mrs. Emerson have?

11. The outcome of a weighted voting mechanism depends not only on the rankings of the individual judges, but also on the points assigned to each place in the rankings. For the beauty contest example described in the text, determine the rankings of the contestants if a second place is worth only 3 points.

12. In some gymnastics competitions, four judges individually assign a number between 1 and 10 to each contestant. The highest and lowest scores are discarded and the contestant receives the average of the two intermediate scores. What injustices would be associated with such a voting mechanism?

13. In many voting situations, the individual voter is permitted to designate more than one contestant as his preference, but is not allowed to rank-order his preferences. For example, 10 candidates may be running for three positions on a local school board. Each voter may place X's besides the names of three candidates. The candidates who receive the largest number of X's are the winners. How fair is such a voting mechanism?

14. Under *plurality* voting, the candidate with the largest number of voters, even if it is not a majority, is declared the winner. Show that with the hypothetical distribution of senators' preferences on tax policy we studied, alternative C would be the plurality winner. Are there legitimate objections to that outcome? Imagine, for example, the 65% of the senators who agreed that B was a better choice than C.

15. *Instant Runoff Voting* (*IRV*) is a procedure that has been gaining substantial support in recent years. If no candidate receives a majority of first-place votes, then the candidate with the fewest first-place votes is eliminated and that candidate's voters are then assigned to the person ranked second on the ballots that named that candidate as first choice. Show that under IRV and hypothetical distribution of senators' preferences on tax policy, alternative A would be eliminated and the 31 votes it received would all be

transferred to alternative B. What inequities does IRV suffer from?

16. Under *Traditional Runoff Voting* (*TRV*), the two candidates with the highest numbers of first-place votes advance to a second election between the pair.

(a) Show that TRV and IRV produce the same result if there are exactly three candidates.

(b) Construct, if possible, a set of rankings among four candidates that yields different winners under TRV and IRV.

III. An Axiomatic Approach

17. Show that simple majority voting satisfies Axioms 1−5 if there are exactly two alternatives.

18. Which of the axioms are satisfied by weighted voting mechanisms?

19. Construct voting mechanisms that satisfy all the axioms except
(a) Axiom 1, (b) Axiom 2, (c) Axiom 3, (d) Axiom 4, (e) Axiom 5

20. Weaken Axiom 2 by eliminating transitivity of societal preferences, and construct various mechanisms that satisfy the new set of axioms.

21. Weaken Axiom 2 by eliminating the demand for a unique societal preference, and construct a mechanism that satisfies the new set of axioms.

22. In what way does Axiom 4 eliminate the possibility of voters manipulating the system by disguising their true preferences?

IV. Arrow's Theorem

23. Construct a voting mechanism for which there is a set V of voters and a pair of alternatives x and y so that V is decisive for x against y, but V is not decisive for y against x. Can you construct such a mechanism that satisfies all but one of Arrow's axioms?

24. The eight-person society discussed under remark (d) has a mechanism that does not satisfy the Unanimity Axiom. Why? Suppose the mechanism is modified so that for every pair x, y of alternatives, x is socially preferred to y whenever everyone prefers x to y; otherwise, the societal preference is the opposite of Mike's. What axioms does this system satisfy?

25. **(a)** Prove that a minimal decisive set will always exist if there is a finite number of voters provided Axiom 2 is satisfied.

(b) Construct a voting mechanism for a society with an infinite number of voters in which minimal decisive sets do not exist—that is, show that if V is any set of voters decisive for some pair of alternatives, then there is a proper subset of V that is also decisive for some pair of alternatives. Can you construct such a mechanism that satisfies all of Arrow's axioms?

26. What happens in the proof of Claim I if the minimal decisive set is the set of all voters? Can this happen in a system satisfying Axioms 1−4?

27. In the proof of Claim I, it is tacitly assumed that V' is non-empty. Can you prove that V' always contains at least one voter?

28. **(a)** Prove that $a\overline{D}b$ implies aDb.

(b) Find an example in which aDb is true, but $a\overline{D}b$ is not.

29. Prove that xDy implies $z\overline{D}x$ and $y\overline{D}x$ if there are exactly three alternatives x, y, and z—that is, verify (5) and (6) of Case 1.

30. Verify the details of the argument of Case 2.

V. Theorem of Gloomy Alternatives

31. Our proof of the Liberal Paradox Theorem assumed that all four of the alternatives w, x, y, and z were distinct from one another. Prove the theorem if this is not true.

32. Some authors reserve the term *Social Decision Function* (*SDF*) for mechanisms that assign to each possible profile of preferences a single top choice or winner rather than a full societal ranking. Suppose we have precisely two voters and two candidates and each voter rank-orders the candidates.

(a) Show that there are four possible profiles and, since there are two possible winners for each profile, show that there are $2^4 = 16$ possible SDFs.

(b) How many possible SDFs are there if there are three voters and two candidates?

33. If there are two voters and three candidates, show there are 36 possible profiles and hence 3^{36} different possible SDFs.

If there are n voters and k candidates, how many different possible SDFs are there?

34. A social decision function is *nondegenerate* or *nontrivial* if for each possible candidate there exists at least one profile of voter preferences under which that candidate is chosen as the winner.

(a) Which of the 16 possible SDFs of Exercise 32 are nondegenerate?

(b) For a situation with exactly two voters and three candidates, exhibit an SDF that is trivial and an SDF that is nondegenerate.

35. Prove the following proposition: Suppose we have a society with two voters and we need to choose a winner among exactly three alternatives using an SDF that is nondegenerate and strategy-proof. If both voters prefer candidate A to candidate B, then B cannot be the winner using this SDF.

36. Use the proposition of Exercise 35 to prove the Gibbard-Satterthwaite Theorem in the case of a two-voter, three-candidate situation. In particular, show that any nondegenerate strategy-proof SDF that returns a winner for every profile of rankings must be dictatorial.

VI. Instant Runoff Voting

37. Table 6.28 below shows some results for the 1990 presidential election in Ireland. The three candidates were Mary Robinson (*MR*) of the Labour Party, Brian Lenihan (*BL*) of the traditionally dominant Fianna Fail Party, and Austin Currie (*AC*) of the Fine Gael Party. A total of 1,584,095 people voted in the election; of these, 9,444 indicated no preference for president. Note that the voters in Group I only listed Robinson as first choice; they did not indicate a second. The voters in Group II listed Robinson first and Currie second, but showed no third choice explicitly.

TABLE 6.28

GROUP	I	II	III	IV	V	VI	VII	VIII	IX
SIZE	306133	183679	122453	208345	347242	138897	205565	46789	25548
1st	*MR*	*MR*	*MR*	*BL*	*BL*	*BL*	*AC*	*AC*	*AC*
2nd		*AC*	*AC*		*AC*	*AC*	*MR*	*BL*	
3rd			*BL*			*MR*			

(a) Who was the plurality winner?

(b) Who is the winner under Instant Runoff Voting?

38. In what sense is the winner of an IRV procedure dependent on which candidate is eliminated first?

39. There are at least two ways that IRV can be modified to produce a societal preference ranking among all the candidates after determining the first-place finisher: (I) Give second place to the last candidate eliminated, third place to the next to last eliminated, and so forth, or (II) Go back to the original individual rankings, cross out the winner, and apply IRV to the result; the winner of this election is the second-place finisher. Continue similarly to find third place, fourth place, and so on.

(a) Apply (I) and (II) to the example shown in Table 6.11.

(b) Apply (I) and (II) to the example in Exercise 37.

(c) Do (I) and (II) always produce the same societal ranking?

(d) Suppose you use method I. Which of Arrow's Axioms are satisfied, and which are violated?

(e) Suppose you use method II. Which of Arrow's Axioms are satisfied, and which are violated?

40. Verify that in the Fishburn-Brams example with candidates Bitt, Huff, and Wogg (Table 6.21), Mr. Huff wins each possible two-person race.

41. For the Fishburn-Brams example with the Smiths voting, show that if two or more of the 82 voters in Group II had switched the order of Bitt and Wogg so that Wogg became their top choice, then Huff would become the winner, again demonstrating that increasing support for a candidate can lead to his defeat!

42. Show that if the Smiths had voted in any order that did not have Mrs. Bitt in first place, then Mr. Huff would have won.

43. Suppose the Smiths not only voted, but also recruited 300 other additional people who shared the Smiths'

preferences but were not intending on voting to show up at the polls and submit their rankings. Who wins the election under IRV?

44. Here is another interesting problem that Fishburn and Brams found can arise with IRV. Suppose the town of Branburn is divided into two districts, East and West, with the numbers of voters in each district with the possible preference rankings summarized in this table:

Group	I	II	III	IV	V	VI
Total	417	82	143	357	285	324
East	160	0	143	0	0	285
West	257	82	0	357	285	39
1st	Bitt	Bitt	Huff	Huff	Wogg	Wogg
2nd	Huff	Wogg	Bitt	Wogg	Bitt	Huff
3rd	Wogg	Huff	Wogg	Bitt	Huff	Bitt

Show that we apply the IRV process to each district separately, then Bitt wins in the East and she wins in the West, but she does not win if the entire town is considered as one district. Fishburn and Brams call this an example of the *Multiple Districts Paradox*: A candidate can win in each district separately, yet lose the general election in the combined districts.

45. Suppose there are nine voters with the following preference rankings for candidates *A*, *B*, and *C*:

Group	I	II	III
Size	4	3	2
1st choice	*A*	*B*	*C*
2nd choice	*C*	*C*	*A*
3rd choice	*B*	*A*	*B*

Show that under plurality or instant runoff voting, *A* wins the election even though *C* is a Condorcet candidate—that is, a majority prefers *C* over *A* and a majority prefers *C* over *B*.

46. Prove that under plurality or instant runoff voting with precisely 3 candidates, a Condorcet candidate will always lose if that candidate is not one of the top two candidates listed in first place. Is it true that a Condorcet candidate will always win if that person is one of the top two? Construct an example, if possible, of a 4 candidate contest with candidates *A*, *B*, *C*, and *D* where *C* is a Condorcet candidate, *C* wins under instant runoff voting, but *A* and *B* each receive more first-place votes than C.

VII. Approval Voting

47. Suppose we have a population of 120 people ranking three alternatives *A*, *B*, and *C*. We can divide the population into five groups whose numbers and orderings of the alternatives are shown in the table below:

Group	I	II	III	IV	V
Size	36	12	26	20	26
1st choice	*A*	*B*	*B*	*C*	*C*
2nd choice	*B*	*C*	*A*	*B*	*A*
3rd choice	*C*	*A*	*C*	*A*	*B*

(a) Show that *C* is the plurality winner.

(b) Show that *A* is the Condorcet Winner.

(c) Show that if first-place votes earn 3 points, second-place votes 2 points, and third-place votes 1 point, then the Borda count winner is *B*.

(d) If we use Approval Voting and each person votes for her top two choices, who wins?

(e) Suppose we use Approval Voting and half of each group votes for their top two choices and the other half only votes for their top choice, who wins?

(f) Who is the winner under Instant Runoff Voting?

48. One argument for Instant Runoff Voting (IRV) over Approval Voting (AV) goes as follows:

Approval voting has another important real-world flaw. Political behavior has much to do with what is rewarded by the election system, and AV would exacerbate one of the worst aspects of U.S. campaigns: avoidance of substantive policy debate. Because a candidate could lose despite being the first choice of an absolute majority of the electorate, smart candidates would avoid controversial issues that alienate any significant number of voters. Smiling more and using policy-empty themes like 'I care' will not clarify the important choices leaders must make. Those rewarded by AV could be characterized as 'inoffensive' more than 'centrist.' IRV strikes a better balance. It rewards

candidates who stand out on policy enough to gain first-choice support, yet encourages coalition-building and fewer personal attacks as candidates seek to be the second choice of other candidates' supporters. [http://archive.fairvote.org/op_eds/science2001.htm]

Discuss the merits of this argument. Could a "smart" candidate in an IRV situation also do well by being "likable" by being "inoffensive" and garner a lot of second choice?

49. Those who favor Approval Voting over Instant Runoff Voting sometimes claim that while it may be "rational" under IRV for a voter to list her second choice as her top choice in the submitted preference ranking, it is never the case under AV that you should vote only for your second choice; you should vote for only your first choice or for your top two choices. Is this claim valid?

50. For our original Senate example with 31 voters having the preference (ABC), 34 having (BCA), and 35 having (CAB), find the winner under Approval Voting if all voters list their top two choices as approved.

51. A voting system is *determinate* if each profile of voters' preference rankings uniquely determines a group preference ranking. One deficiency of Approval Voting is that it is not determinate: given a complete list of everyone's preference rankings, there may be many possible outcomes depending on exactly how many voters approve of more candidates than their first choice. As an example, suppose we have 15 voters whose rankings among candidates A, B, and C divide into three groups with different numbers as shown in this table:

Group	I	II	III
Size	6	5	4
1st choice	*A*	*B*	*C*
2nd choice	*C*	*C*	*B*
3rd choice	*B*	*A*	*A*

(a) Show that if all voters only check their top choice, then candidate A wins.

(b) Show that if exactly two voters of Group III check their top two choices and everyone else only checks their first choice, then candidate B wins.

(c) Show that if exactly three members of Group I check their top two choices and everyone else only checks their first choice, then candidate C wins.

(d) What is the result if four members of Group I and three members of Group III check their top two choices?

(e) What is the result if two members of Group I and one member of Group III check their top two choices?

VIII. Topological Choice Theory

52. Dave and Judy are deciding between two possible locations A and B. They agree on a decision rule that will select location A unless both of them name B as their top choice, in which case B is declared the winner. Does this rule satisfy the properties of anonymity, unanimity, and stability?

53. Suppose Dave and Judy are trying to decide among three possible locations A, B, C along the lakeshore. A proposed decision rule is to choose the common site if they both agree, but to pick the unnamed site if they disagree. Thus, if Dave submits A and Judy submits C, the decision rule outputs B. Show that this rule satisfies the properties of anonymity, unanimity, and stability.

54. Show that the assignment shown in the matrix below violates the stability criterion.

		DAVE'S A	TOP B	CHOICE C	D
JUDY'S	A	*A*	*A*	*B*	*D*
TOP	B	*A*	*B*	*C*	*D*
CHOICE	C	*B*	*C*	*C*	*C*
	D	*D*	*D*	*C*	*D*

55. For the David and Judy problem with five possible home sites, find social choices that are:

(i) Unanimous and anonymous but not stable

(ii) Unanimous and stable but not anonymous

(iii) Stable and anonymous, but not unanimous

56. A *metric* on a set S is a real-valued function d that assigns a nonnegative number $d(x, y)$ to every pair of elements x and y of S in such a manner that for all x, y, and z in S, the following are true:

(i) $d(x, y) = 0$ if and only if $x = y$

(ii) $d(x, y) = d(y, x)$

(iii) $d(x, z) \leq d(x, y) + d(y, z)$ (triangle inequality)

(a) If S is the set of real numbers, show that the function $d(x, y) = |x - y|$ is a metric.

(b) If S is the set of points in the plane, show that the function $d(x,y)$ = length of line segment between x and y is a metric. Note that if $x=(x_1,x_2)$ and $y=(y_1,y_2)$, then $d(\mathbf{x},\mathbf{y})=\sqrt{(x_1-y_1)^2+(x_2-y_2)^2}$ is a metric.

(c) If S is the set of all n-dimensional vectors $\mathbf{x}=(x_1, x_2,\ldots,x_n)$ where each x_i is a real number, show that $d(\mathbf{x},\mathbf{y})=\sqrt{(x_1-y_1)^2+(x_2-y_2)^2+\cdots+(x_n-y_n)^2}=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2}$ is a metric. This metric is called the *Euclidean metric* or Euclidean distance function.

(d) If $n=1$, how do the metrics in parts (c) and (a) compare?

57. Let $x=(x_1, x_2)$ and $y=(y_1, y_2)$ be any two points in the plane. Show that the function d defined by $d(x,y)=|x_1-y_1|+|x_2-y_2|$ is a metric. Can you say why this function d is called the *taxicab* metric?

58. Show that the $\varepsilon-\delta$ definition of continuity is equivalent to the *neighborhood* definition.

59. **(a)** The *norm* of a vector $\mathbf{v}=(v_1,v_2,\ldots,v_n)$ is the real number $||\mathbf{v}||=\sqrt{v_1^2+v_2^2+\ldots+v_n^2}$. Show that the norm of a vector is the Euclidean distance between the origin $0=(0,0,\ldots,0)$ and the vector.

(b) If $\mathbf{v}$ is a nonzero vector, then show that $\mathbf{u}=\frac{\mathbf{v}}{||\mathbf{v}||}$ is a vector of length 1 pointing in the same direction as $\mathbf{v}$.

60. Show that there is no social choice rule for the David and Judy problem with five possible home sites that is unanimous, anonymous, and stable.

61. Let X be the set of points on the unit circle except for the point $(-1, 0)$. Show that each point P on X can be described uniquely by an angle θ where $-\pi<\theta<\pi$. Then consider the function F that sends the pair (θ,t) to the point $(1-t)q$ for $0\le t\le 1$ Show that F sends each point of the form $(\theta,0)$ to itself and each point of the form $(\theta,1)$ to the point with Cartesian coordinates $(1, 0)$. Finally, show that F is a continuous function of θ and t. Thus the effect of F is to continuously shrink X to a point in X, always staying inside X. Such a function is called a *retraction*. A space is *contractible* if there is a retraction of X onto a single point in X. Thus, the circle minus a point is contractible.

62. Suppose our city council initially wishes to divide its expenditures into two categories: Municipal Services (M) and Education (E). The council will spend at least \$10 million, but no more than \$20 million.

(a) Sketch the set of all possible choices.

(b) Sketch the set of all possible choices if each of M and E must be at least \$5 million.

63. Are the regions in Exercise 62 (a) and (b) contractible?

64. The city council needs to pick a site for a new water treatment plant. It decides that the plant must be at least 2 miles from the center of the city, but no more than 3 miles from the center. Sketch the set of all possible locations. Is this set contractible?

65. Suppose the preference rankings among nine voters are distributed as in Exercise 45.

(a) Show that if 10 points are awarded for each first-place ranking, 5 points for second, and 2 points for third, then A wins a Borda count even though C is the Condorcet Winner.

(b) If N is any positive integer and N points are given for first-place rankings, $N-1$ for second and $N-2$ for third, then C wins under a Borda count.

66. Construct an example of preference rankings where a Borda count with N points for first place, $N-1$ for second, $N-2$ for third, and so forth yields a winner who is not the Condorcet Winner.

67. Consider the preference rankings given in Table 6.27.

(a) Show that if any three voters belonging to Groups I–IV interchange their consecutive rankings of C and A, then A becomes a Condorcet Winner.

(b) What is the smallest number of interchanges of consecutive candidates needed to make C the Condorcet Winner? Answer the same question for candidates B and D.

SUGGESTED PROJECTS

1. Instead of discarding some axioms or weakening them to obtain a consistent set, you might think about strengthening the first axiom. Axiom 1 allows voters to list the alternatives in order of preference, but does not allow for expression of intensity of differences between alternatives. Two voters may both list x and y

at the top of their lists, although the first voter's feelings are almost indifferent toward the two, while the second voter much prefers x to y. Investigate methods of incorporating intensities into individual preference lists. Discuss the consistency of sets of axioms allowing for such measures.

One such approach is *Range Voting*, a system whereby each voter assigns a measure on a scale of 0 to 100 to each candidate where the score reflects the strengths of the voter's preference for that candidate or voter's rating of the candidate's worth or appeal to the voter. The scores are added; the candidate with the highest total score is the winner. Range Voting can be interpreted as a generalization of approval voting that uses a scale of 0 (disapprove) to 1 (approve). Show that we can derive each voter's preference ranking of the candidates from the scores given out. What are the strengths and weaknesses of Range Voting?

2. Peter Fishburn has shown that the Axioms 1−5 are consistent if there are an *infinite* number of voters. Investigate his proof. What is the real-world relevance of this result?

3. Some voting theorists have argued that the modified simple majority vote system is satisfactory because intransitivity rarely occurs. Is there some way of measuring the likelihood of intransitivity? Can you find instances in Senate voting where proponents have used intransitivity to their advantage by adjusting the agenda?

4. Can the standard voting systems (simple majority, weighted voting, and so on) be characterized axiomatically? H. P. Young isolated three characteristics of voting systems, which he termed "consistency," "the cancellation property," and "faithfulness." He was able to prove that any mechanism that is consistent, faithful, and has the cancellation property must be a weighted voting system. Are these three properties reasonable ones? Check the details of Young's proof. Derive, if possible, an axiomatic characterization of simple majority voting.

5. Develop a proof of Arrow's Theorem showing that Axioms 1, 2, 4, and 5 imply that Unanimity is violated.

6. Show how the proof for the Muller-Satterthwaite Theorem can be easily modified to prove Arrow's Theorem.

7. A *Condorcet method* is any election method that always selects the Condorcet Winner, the candidate who would beat each of the other candidates in a run-off election, if such a candidate exists. Which of the social choice rules we have studied are Condorcet methods? Arthur Copeland suggested one of the simplest Condorcet methods in 1951: pick the candidate who beats the largest number of other candidates in run-off elections. Investigate the properties of the Copeland method.

8. Perhaps the most current widely used Condorcet method is a process suggested by Markus Schulze in 1997. Although the Schulze method (also known as *Path Voting*) fails to satisfy the Independence of Irrelevant Alternatives Axiom, it does satisfy a number of other desirable criteria for a fair voting mechanism. Determine how the Schulze method works, what societal rankings it outputs for the examples of profiles we have see in this chapter, and proofs that it satisfies some of the fairness criteria and fails others. Begin with Schulze's paper "A New Monotonic and Clone-Independent Single-Winner Election Method," *Voting Matters* **17** (2003): 9−19. Schulze expands on this work in a number of papers available on his website http://m-schulze.webhop.net

9. *Computational Social Choice* is a new field that explores questions at the interface of social choice theory and computer science. One of its main concerns is the efficiency of algorithms to implement the variously suggested voting mechanisms. How rapidly, for example, does the number of computational steps increase for a particular mechanism as the number of individual voters or individual candidates increase? Which mechanisms that appear attractive in theory may be infeasible to implement in practice because it would take the fastest computers hundreds of years to determine the winner? Are there social decision procedures that in theory are not strategy-proof but that in practice pose a computationally intractable problem to determine how to manipulate them by insincere rankings? Investigate what is known about the efficiency of particular social choice procedures. See Chevaleyre et al. (2005) for a good introduction to this interdisciplinary field.

10. Manipulable voting processes pose a serious problem for many social choice theorists as they may encourage voters to disguise their true preferences when submitting their rankings of the candidates. Others see that manipulation can take different forms and some may be a healthy aspect of democracy. Keith Dowding and Martin Van Hees distinguish between *sincere* and *insincere* manipulation and examine a class of social choice functions that are immune to one or the other

form. Investigate the claims they make in their provocatively titled paper "In Praise of Manipulation," *British Journal of Political Science* **38** (2007): 1–15 and examine what they imply about the main voting processes we have discussed.

11. Although Instant Runoff Voting and Approval Voting appear now to be the leading candidates to replace current methods of collective decision making, there are also strong arguments for using the Borda count. Examine Donald Saari's (2001) arguments on this option.

12. The Coombs Rule is an interesting variant of Instant Runoff Voting. This rule, suggested by mathematical psychologist Clyde Coombs, eliminates the candidate with the most last-place votes rather than the one with the fewest first-place votes. Investigate the properties of the Coombs Rule. In what ways is it superior to IRV? A good starting reference is Bernard Grofman and Scott L. Feld, "If You Like the Alternative Vote (a.k.a the Instant Runoff), Then You Ought to Know about the Coombs Rule," *Electoral Studies* **23** (2004): 641–659.

13. Many of the issues, definitions and techniques that arise in social choice theory are related to issues of *Fair Division* and *Apportionment*. Fair Division, also known as the *cake-cutting problem*, is the problem of dividing a resource in such a way that all recipients believe that they have received a fair amount. An important apportionment problem is allocating the 435 seats in the U.S. House of Representatives among the states in proportion to their populations with the constraints that each state gets a whole number of representatives. Examine the various methods proposed to solve such problems. A good starting reference is H. Peyton Young, *Equity in Theory and Practice*, Princeton: University Press, 1994.

14. Students with a background in algebraic topology would enjoy preparing an exposition of some of the results of Chichilnisky, Heal, and Baryshnikov listed below in the References.

15. Investigate social decision procedures that combine preference rankings and approval voting. In such systems, a voter either (a) submits a ranking of all the candidates, drawing a line between those approved by the voter and those not approved, or (b) submits a ranking only of those candidates approved by the voter. Various rules for determining a winner can be implemented, giving rise to different fairness criteria being satisfied. See Steven J. Brams and M. Remzi Sanver, "Voting Systems That Combine Approval and Preference," in Steven Brams, William Gehrlein and Fred Roberts, eds., *The Mathematics of Preference, Choice and Order: Essays in Honor of Peter C. Fishburn*, Berlin: Springer-Verlag, 2009, 215–237.

16. The Scottish economist Duncan Black (1908–1991) argued that modified simple majority voting is a fair social decision procedure if certain conditions on individual preferences prevented intransitive results from occurring. Examine Black's notions of *single peaked preferences* and *median voter* and his formulation and proofs of some possibility theorems. See Duncan Black, "On the Rationale of Group Decision Making," *Journal of Political Economy* **56** (1948): 23–34; and A. K. Sen and P. K. Pattanaik, "Necessary and Sufficient Conditions for Rational Choice under Majority Decision," *Journal of Economic Theory* **1** (1969): 178–202.

You can find a listing of references and suggestions for additional reading on the book's website, www.wiley.com/college/olinick

CHAPTER 7 Foundations of Measurement Theory

It is a scientific platitude that there can be neither precise control nor prediction of phenomena without measurement. Disciplines as diverse as cosmology and social psychology provide evidence that it is nearly useless to have an exactly formulated quantitative theory if empirically feasible methods of measurement cannot be developed for substantial portions of the quantitative concepts of the theory.

—Dana Scott and Patrick Suppes

I. The Registrar's Problem

Middlebury College divides its academic year into three major components: two 12-week semesters sandwiching a 4-week "winter term." During the winter term each faculty member offers, and each student enrolls in, one course. Because of the experimental nature of many of the courses offered, enrollment is often restricted to 20 or 25 students in each class. Since a typical Winter Term will find 1,800 students on campus and only 70 courses, it is clear that not every student will be able to take the course she most desires.

When a student registers for winter term, then, she lists five courses in descending order of preference. The registrar assigns each student to a course, using these preferences as a guide. At the present time, the registrar uses a procedure based on the desire to maximize the number of students who receive their first choice. There has been considerable discussion lately about the fairness and desirability of this particular priority scheme. An alternative method of assigning students to courses has been devised, which has gained some support. The philosophy behind this scheme is not to maximize the number of first choices, but to maximize the total amount of happiness among the students towards the courses they are assigned. This assignment procedure can be given a rather tidy mathematical formulation.

Denote the students by $i = 1, 2, \ldots, n$ and the courses by $j = 1, 2, \ldots, m$. Let r_{ij} denote how happy student i would be if she is assigned course j. Define the variable x_{ij} to be equal to 1 if student i is placed in course j and 0 otherwise. The total amount of happiness would then be represented by

$$r_{11}x_{11} + r_{21}x_{21} + \cdots + r_{n1}x_{n1} + r_{12}x_{12} + \cdots + r_{nm}x_{nm}$$

or, in more compact form,

$$\sum_{j=1}^{m}\sum_{i=1}^{n} r_{ij}x_{ij}.$$

There are several restrictions on the registrar that must be taken into account. First of all, every student must be assigned to some course and only to one course. Second, no course should be assigned more students than the instructor is willing to admit. Denoting the enrollment limit on the jth course by C_j, the Registrar's Problem is formulated as follows:

$$\text{Maximize} \quad \sum_{j=1}^{m}\sum_{i=1}^{n} r_{ij}x_{ij}.$$

subject to the following restrictions:

1. Each $x_{ij} \geq 0$.
2. $\sum_{j=1}^{m} x_{ij} = 1$ for all i.
3. $\sum_{i=1}^{n} x_{ij} \leq C_j$ for all j.

The Registrar's Problem is an example of what mathematicians call a "linear programming" problem: optimizing a linear function subject to a set of linear constraints. Algorithms exist for the solution of such linear-programming problems although when n and m are as large as 1,800 and 70, respectively, digital computers must be used to execute them.

In this chapter and the succeeding one, we want to focus more sharply on the aspect of the Registrar's Problem, which remains somewhat vague in this presentation: what precisely is r_{ij} and how is it determined?

II. What Is Measurement?

A. A Physical Analogy

We let r_{ij} denote "how happy student i would be if she is assigned course j." In the mathematical formulation of the Registrar's Problem, it is clear that we are presuming that each r_{ij} is a real number that measures this happiness. Is it clear, however, that it is always possible to measure such psychological attributes by numbers? What is meant by "measuring" an attribute? Is there more than one way to do it? What inferences, if any, can be made from a measurement scale? How can you construct such a scale?

These questions form the basic problems of *measurement theory*. In the mid-1960s four distinguished social scientists [David Krantz, R. Duncan Luce, Patrick Suppes, and Amos Tversky, 1972] began a collaborative study of the foundations of measurement theory that resulted in three large volumes. Early in their first book, the authors discuss the roles of theories of measurement in science:

> *The measurability of the variables of interest in physics is taken for granted and the actual measurements are reduced, via the elaborate superstructure of physical theory, to comparatively*

> *indirect observations. Other sciences, especially those having to do with human beings, approach measurement with considerably less confidence. In the behavioral and social sciences we are not entirely certain which variables can be measured nor which theories really apply to those we believe to be measurable; and we do not have a superstructure of well-established theory that can be used to devise practical schemes of measurement. . . . A recurrent temptation when we need to measure an attribute of interest is to try to avoid the difficult theoretical and empirical issues posed by fundamental measurement by substituting some easily measured physical quantity that is believed to be strongly correlated with the attribute in question: hours of deprivation in lieu of hunger; skin resistance in lieu of anxiety; milliamperes of current in lieu of aversiveness, etc.*

It should not be surprising then that our first insights into measurement will come from considerations of measurement in the physical sciences. The question of what is meant by "measuring an attribute" may perhaps best be answered by examining first a physical attribute, weight. A provisional definition of *measuring an object's weight* might be "assign some number to that object." This is a very poor definition, since the same number might be assigned to every object.

A careful analysis of a physical attribute is not possible unless there is some means of deciding which of two objects possesses more of the attribute than the other. A refinement of the first definition might be the following: to measure an object's weight means to assign a number to that object in such a way that one object is at least as heavy as a second object if and only if it is assigned a number at least as large as the second.

This added restriction rules out the possibility of assigning all objects the same number. It relies on the fact that the concept of the "weight" of an object is intimately connected with a relation between objects, the relation "at least as heavy as." This relation can be established empirically by placing any pair of objects on the separate pans of a balance and observing which pan descends.

Write A^*B if object A is at least as heavy as object B. To measure weight is to find a function w from the set of objects to the set of real numbers such that $w(A) \geq w(B)$ if and only if A^*B.

It is natural to define "A has the same weight as B" to mean that A^*B and B^*A. As an easy exercise, the reader should show that $w(A) = w(B)$ if and only if A has the same weight as B.

It is now easy to describe a procedure for assigning weights to a finite set of objects $A_1, A_2, \ldots, A_n$. By testing A_1 against each of the other objects on the pan balance, then A_2 against all the other objects, and so on, find a lightest element A_j. This is an object A_j such that $A_i{}^*A_j$ for all i. Assign weight 0 to object A_j. If there is any A_i such that $A_j{}^*A_i$ then A_j and A_i have the same weight, so also assign weight 0 to A_i.

Next consider the set of remaining objects that have yet to be assigned a weight. Find a lightest element in this set. Assign weight 1 to this object and to any object of the same weight. Repeat the process on the set of remaining objects (assigning weight 2 to its lightest element) and continue in this manner until all objects have been assigned a weight.

B. Relations

With this relatively simple example as background, we can discuss the general problem of defining what it means to measure attributes. The formulation of the problem uses the concept of a "relation" from elementary set theory. See Appendix I for the necessary background on sets.

DEFINITION A *relation* on a set S is a subset R of the Cartesian product $S \times S$. If x and y are elements of S, we say that *x is R-related to y* or xRy whenever (x, y) is an element of R.

We will present a number of examples to illustrate this idea.

Example 1

Let S be a set with four elements, $S = \{a, b, c, d\}$. The Cartesian product $S \times S$ consists of 16 ordered pairs:

$$S \times S = \{(a, a), \ldots, (a, d), (b, a), \ldots, (d, a), \ldots, (d, d)\}$$

A relation on S consists of some subset of these sixteen ordered pairs. One such example is a relation with three elements, $R = \{(a, c), (a, d), (b, d)\}$. We have aRc, aRd, and bRd, and for no other pair i and j is it true that iRj.

Example 2

Let S be the set of all positive integers and consider the relation R defined by xRy if and only if the difference $x - y$ is even. Thus, (2, 4) is an element of R, while (3, 2) is not. The relation R consists of all pairs (x, y) such that either both x and y are even or both x and y are odd.

Example 3

Let S be the set of all real numbers and let R be the set of all ordered pairs (x, y) such that $x \geq y$. Note that $(5, 3)$ is an element of R, but $(3, 5)$ is not. Since the Cartesian product $S \times S$ consists of all ordered pairs of real numbers, it can be represented geometrically by the points in the plane. Any relation on S then corresponds to some subset of the plane. Fig. 7.1 is a graphical representation of this relation R.

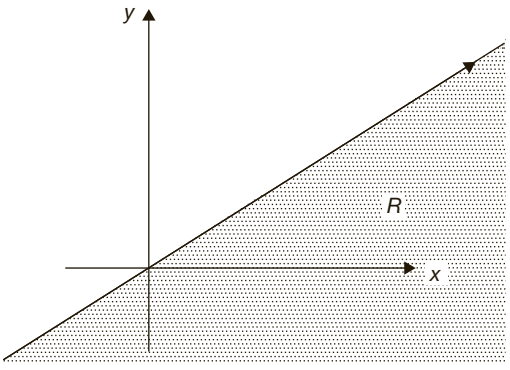


FIGURE 7.1 The shaded region R consists of all pairs (x, y) of real numbers such that $x \geq y$.

Example 4

Let S be the set of all people in Georgia, and let R be the relation defined by xRy if and only if x knows y.

Example 5

Let S be the set of all people in the U.S. Navy, and let R be the relation defined by xRy if and only if y must obey an order given by x.

Example 6

Let S be the set of all automobiles in Honest Harry's Used Car Lot. Define a relation xRy if and only if x costs more than y.

Example 7

Let S be the set of all objects in your attic, and let R be the relation defined by xRy if and only if x is at least as old as y.

Example 8

Let S be the set of all words in the English language, and let R be the relation defined by xRy if and only if x precedes y in the dictionary.

Example 9

Let S be the set of all ordered pairs of real numbers. Define a relation $(x, y)R(x', y')$ if and only if $x < x'$ or ($x = x'$ and $y < y'$). For example, we have $(3, 20)R(5, 11)$ and $(3, 20)R(3, 21)$. This relation is called the *lexicographic* or *dictionary order*.

Example 10

Let S be the set of all courses offered by your college, and let R be the relation xRy if and only if you like course x at least as much as course y.

Example 11

Let S be the set of all courses offered by your college, and let R be the relation xRy if and only if course x is a prerequisite for course y.

Scientists classify relations by the presence or absence of certain properties. We will consider here a few of the more important properties.

DEFINITION If S is a set and R is a relation on S, then

1. R is *reflexive* if xRx for all x in S
2. R is *symmetric* if xRy always implies yRx
3. R is *transitive if* xRy and yRz always implies xRz
4. R is *connected* or *total* if for every pair of elements x and y in S, either xRy or yRx or both

Example 3 is reflexive and transitive, but not symmetric. Example 2 is symmetric. Example 6 is transitive, but neither reflexive or symmetric. Example 4 is probably not transitive. Examples 3 and 7 are connected, while Examples 1 and 2 are not.

C. Definition of Measurement

This section provides a careful definition of measurement and explores some of its elementary consequences. By a *relational system*, we mean a pair $\alpha = < S, R >$ where S is a set and R is a relation on S. A *measure* for a relational system α is a function m from S to the real numbers such that for all x and y in S, xRy if and only if $m(x) \geq m(y)$.

To measure an attribute possessed by a set of objects, people, or events means to find a measure m that preserves the relation determined by the attribute. The "Basic Representation Problem" then is: Which relational systems have measures?

Note first that it is not always possible to find a measure for a given relational system.

Example 12

Let S be the set of three elements $\{x, y, z\}$ and let R be the relation $\{(x, y), (y, z), (z, x)\}$. The relational system $< S, R >$ has no measure. Suppose, to the contrary, that there is a measure m. Since xRy and yRz, we must have $m(x) \geq m(y)$ and $m(y) \geq m(z)$. But $m(x)$, $m(y)$, and $m(z)$ are real numbers so it follows that $m(x) \geq m(z)$. Since m is a measure, the definition implies that xRz or (x, z) is an element of R. The ordered pair (x, z), however, does not belong to R. The assumption that $< S, R >$ has a measure that leads to a contradiction.

The relational system described in Example 12 failed to have a measure essentially because the relation was not transitive. The reasoning given in the discussion of this example extends to a more general situation, stated in Theorem 1. We leave the proof as an exercise.

THEOREM 1 If a relational system $\alpha = < S, R >$ has a measure, then R is a transitive relation.

Theorem 1 says that one necessary condition for a relational system to have a measure is that the relation be transitive. It is easy to establish a second necessary condition—namely, the relation must be connected.

THEOREM 2 If the relational system $\alpha = < S, R >$ has a measure, then R is a connected relation.

Proof If x and y are any two elements of S, then $m(x)$ and $m(y)$ are defined and are real numbers. It must be true that either $m(x) \geq m(y)$ or $m(y) \geq m(x)$. In the former case, xRy and in the latter, yRx. Thus, either $(x, y) \varepsilon R$ or $(y, x) \varepsilon R$. ◇

Theorems 1 and 2 indicate that the relational systems of Examples 1 and 2 have no measures associated with them.

III. Simple Measures on Finite Sets

One of the major goals of measurement theory is to establish necessary and sufficient conditions on relational systems under which various numerical representations can be constructed. The relation must be connected and transitive if there is to be any hope of constructing a measure. If the set S is finite, then these two conditions are also sufficient.

THEOREM 3 (FIRST REPRESENTATION THEOREM) Let R be a relation on a finite set S. There exists a measure on the relational system $< S, R >$ if and only if R is connected and transitive.

Proof Theorems 1 and 2 establish the "only if" part of the conclusion. It remains to show that if R is connected and transitive, then it is always possible to find a measure. The idea behind the proof is essentially the same as the one used in describing how to assign numerical weights to a set of objects.

Denote the elements of the set S by $x_1, x_2, \ldots, x_n$. Since the relation R is connected and transitive, we can find, by checking all possible pairs of elements, an element x_j such that $x_i R x_j$ for all $i \neq j$. Define $m(x)$ to be 0. If there is any element x_i so that $x_j R x_i$ as well as $x_i R x_j$, then define $m(x_j)$ to be 0 also.

At this point, at least one and possibly more elements of S have been assigned measure 0. Consider the subset S' of remaining elements. Find an element x_k of S' so that $x_m R x_k$ for all $x_m \neq x_k$ in S'. Define $m(x_k) = 1$. If there is any other element x_m of S' with $x_k R x_m$ as well as $x_m R x_k$, then also define $m(x_m) = 1$.

Repeat the entire process on the set S'' of elements that have not yet been assigned measures, using a measure of 2 to distinguish one or more special members of S''. Continue in the indicated manner until each element of S has been assigned a measure. This will take at most n steps.

It should be clear from the method of construction that m satisfies the definition of a measure—that is, xRy if and only if $m(x) \geq m(y)$. This completes an outline of a proof of Theorem 3. ◇

We used the finiteness of the set S at several crucial steps in the proof of Theorem 3. The theorem remains true if S is a countably infinite set, but may fail if S is uncountable; see the Exercises for the relevant definitions and examples.

Note that the proof of Theorem 3 not only establishes the existence of a measure, but provides an effective method of constructing one.

The numerical values of a measure function m are sometimes called *scale values*. In the proof of Theorem 3, the numbers 0, 1, 2, and so on were suggested for scale values. There is nothing sacred about this set. We could have used any increasing sequence of real numbers. The next example amplifies this point.

Example 13

Let S be the set of three elements $\{x, y, z\}$ and R the relation $\{(x, y), (y, z), (x, z)\}$. Since this relation is connected and transitive, the elements can be represented numerically by a measure, according to Theorem 3. If the procedure outlined in the proof of that theorem is followed, the result is

$$m(z) = 0, \quad m(y) = 1, \quad \text{and} \quad m(x) = 2.$$

These values are not determined by the measurement model. We could set

$$m(z) = -17, \quad m(y) = \sqrt{23}, \quad \text{and} \quad m(x) = 10$$

and still satisfy the definition of a measure. In fact, any three numbers $m(x)$, $m(y)$, $m(z)$ satisfying the inequalities $m(z) < m(y) < m(x)$ would be an admissible set of scale values.

Such scales are called *ordinal scales*. Any transformation of the scale numbers that preserves their original order yields another admissible scale. A transformation that changes the order in any way would give a set of scale values that is not admissible. The resulting numbers, $m(x)$, would define a function that is not a measure.

If $\alpha = < S, R >$ is a relational system where S is finite and R is connected and transitive, then the elements of S can be labeled $x_1, x_2, \ldots, x_n$ in such a way that if m is any measure, then $m(x_1) \leq m(x_2) \leq \cdots \leq m(x_n)$. This is called the *standard ordering* on S.

The mathematical model just developed gives a partial solution to the question posed by the Registrar's Problem. It is possible to assign numbers to a given student that measure her happiness about being enrolled in the available courses exactly when the student can state her relative preference for each pair of courses, provided these preference judgments are transitive.

By asking the student a series of questions requiring her always to indicate which one of two courses she prefers to the other, we can construct her preference ordering among all 70 courses.

The reason this model gives only a partial solution to the question originally asked will become apparent in the next section of this chapter.

IV. Perception of Differences

Suppose the measure guaranteed by Theorem 3 is used to determine numerical values r_{ij} measuring the relative happiness of students toward courses. We soon encounter a student who complains, "I would be almost equally happy with my first as with my second-choice course, but quite a bit less happy with the third choice. If you assign measures of 3, 2, and 1 to these choices, you are not really representing my feelings in a completely accurate way."

This type of objection forces us to ask if it is possible to choose a scale of numbers that will accurately reflect the differences perceived by the student between different *pairs* of courses. Let's consider a mathematical formulation of this question.

Write $(x, y)\, R^*\, (z, w)$ to denote the student's judgment that the difference in happiness between courses z and w does not exceed the difference in happiness between courses x and y. Note that R^* defines a relation on the set $S \times S$. This type of relation, which is a subset of the set $(S \times S) \times (S \times S)$, is called a *quaternary relation* on S as opposed to a *binary relation*, which is a subset of $S \times S$.

The problem is to find a measure that preserves both R and R^*. More precisely, does there exist a real-valued function u defined on the set S such that for all x, y, z, w in S,

1. $u(x) \geq u(y)$ if and only if xRy, and
2. $u(x) - u(y) \geq u(z) - u(w)$ if and only if $(x, y)R^*(z, w)$?

If u is any real-valued function defined on S, then u induces a connected, transitive relation on S. Simply define a relation R' by $xR'y$ if and only if $u(x) \geq u(y)$. The question can then be posed this way: Is there a real valued function u on S that preserves R^* such that the induced relation R' is identical to the relation R?

Consider first the simpler question: Is there a real-valued function on S that preserves the relation R^*? Using reasoning similar to that in the proofs of Theorems 1 and 2, one concludes that an affirmative answer can be expected only when R^* is connected and transitive. The exact result is stated in the next theorem.

THEOREM 4 Suppose S is a set and R^* is a quaternary relation on S. If there is a real-valued function u defined on S such that

$$u(x) - u(y) \geq u(z) - u(w) \quad \text{if and only if} \quad (x, y) \quad R^*(z, w)$$

then the relation R^* satisfies four properties:

1. R^* is connected
2. R^* is transitive
3. If $(x, y)R^*(z, w)$, then $(x, z)R^*(y, w)$
4. If $(x, y)R^*(z, w)$, then $(w, z)R^*(y, x)$

Proof If $x, y, z,$ and w are any elements of S, then the real numbers $u(x) - u(y) = A$ and $u(z) - u(w) = B$ must satisfy the inequality $A \geq B$ or the inequality $B \geq A$. In the former case, $(x, y)\, R^*\, (z, w)$, and in the latter, $(z, w)\, R^*\, (x, y)$. Thus, R^* is connected.

Next, suppose $(x, y)\, R^* (z, w)$ and $(z, w)\, R^* (a, b)$, where x, y, z, w, a, b are arbitrary elements of S. We have the inequalities

$$u(x) - u(y) \geq u(z) - u(w)$$

and

$$u(z) - u(w) \geq u(a) - u(b)$$

which imply by transitivity

$$u(x) - u(y) \geq u(a) - u(b)$$

so that $(x, y)\, R^* (a, b)$. This shows that R^* is transitive.

Condition (3) is satisfied, since if $(x, y)\, R^* (z, w)$, then

$$u(x) - u(y) \geq u(z) - u(w)$$

which implies (by adding like terms to each side of the inequality)

$$u(x) - u(y) - u(z) + u(y) \geq u(z) - u(w) - u(z) + u(y)$$

or

$$u(x) - u(z) \geq u(y) - u(w).$$

This inequality, in turn, gives $(x, z)\, R^* (y, w)$ by the hypothesis on u.
The proof that condition (4) holds is left to the reader. ◇

The four conditions of Theorem 4 are necessary for the existence of the required measure u, but unlike the case for binary relations, they turn out not to be sufficient. There exists a finite set and a quaternary relation R^* on it that satisfies the four conditions but for which it is not possible to construct a numerical scale preserving R^*.

In a 1958 paper in the *Journal of Symbolic Logic*, Dana Scott and Patrick Suppes prove an even stronger result: if S is a finite set and R^* is a quaternary relation on S, then there is no finite list of axioms that provides necessary and sufficient conditions for the existence of a real-valued function u preserving R^*.

Scott and Suppes cite an example, essentially drawn from Herman Rubin, that indicates the kind of difficulty that arises in trying to construct a set of necessary and sufficient conditions.

Example 14

A student is presented a list of 10 possible courses. By comparing each of the courses with every other one, her order of preference is determined to be $x_1, x_2, \ldots, x_{10}$. Eleven pairs of courses are given special designations: Denote

(x_1, x_2) by A	(x_7, x_8) by E	(x_5, x_6) by I
(x_2, x_3) by B	(x_9, x_{10}) by F	$(x_1 x_5)$ by J
(x_3, x_4) by C	(x_6, x_7) by G	(x_6, x_{10}) by K
(x_4, x_5) by D	(x_8, x_9) by H	

In each pair, the first course is preferred to the second. Suppose the student perceives A, B, C, D as equal in difference to E, F, G, H, respectively, that the difference between courses in K is greater than the difference in courses in J, and that the difference in I is greater than the difference in K. Then the relations between the remaining pairs may be chosen so that any subset of nine courses can be represented by a measure u that preserves R^*, but the full set of 10 courses cannot! The interested reader may wish to work out the details of this example.

V. An Alternative Approach

Since it is not possible to discover or prove a Representation Theorem for arbitrary finite sets and quaternary relations, we need to try some alternative approaches to the problem of measurement that will satisfy our complaining student of Section IV. We will present one alternative in this section and another, called *Utility Theory*, in Chapter 8.

We obtain the relation R by asking a subject to compare each pair of elements in a set and to give her judgment on which possesses more of the relevant attribute than the other. We then obtain the relation R^* by asking the subject to compare each pair of elements with every other pair. We can derive a measure u if we restrict ourselves to asking for comparisons only between pairs when the pairs represent elements that the subject perceives as being consecutive elements in the ordering. If the student has ranked the courses in the order $x_1, x_2, x_3, \ldots, x_n$, then we ask for comparisons when the pairs are $(x_1, x_2), (x_2, x_3), (x_3, x_4), \ldots, (x_{n-1}, x_n)$. Theorem 5 states more carefully the alternative approach based on this idea.

THEOREM 5 (SECOND REPRESENTATION THEOREM) Let $\alpha = < S, R >$ be an ordered relational system where $S = \{x_1, x_2, x_3, \ldots, x_n\}$ is the standard ordering on S. Let T be the set of all ordered pairs (x_i, x_j) where $j = i + 1$ and let $R^{\S}$ be a relation on T. Then there is a measure u on α satisfying the following two conditions:

1. $u(x_i) \geq u(x_j)$ if and only if x_i, Rx_j.
2. $u(x_{i+1}) - u(x_i,) \geq u(x_{j+1}) - u(x_j)$ if and only if $(x_i, x_{i+1}) R^{\S} (x_j, x_{j+1})$ exactly when $R^{\S}$ is connected and transitive.

Proof of Theorem 5 We outline the proof of sufficiency. Since T is finite and $R^{\S}$ is connected and transitive, there is a positive-valued measure $m^{\S}$ for the system $< T, R^{\S} >$. More precisely, $m^{\S}$ is a function from T to the positive real numbers such that

$$m^{\S}(x_i, x_{i+1}) \geq m^{\S}(x_j, x_{j+1}) \text{ if and only } (x_i, x_{i+1}) R^{\S} (x_j, x_{j+1})$$

We can then define a measure u as follows:

$$\begin{aligned}
&\text{Let } u(x_1) = 0\\
&u(x_2) = u(x_1) + m^{\$}(x_1, x_2) = m^{\$}(x_1, x_2)\\
&u(x_3) = u(x_2) + m^{\$}(x_2, x_3) = m^{\$}(x_1, x_2) + m^{\$}(x_2, x_3)\\
&u(x_4) = u(x_3) + m^{\$}(x_3, x_4) = m^{\$}(x_1, x_2) + m^{\$}(x_2, x_3) + m^{\$}(x_3, x_4)\\
&u(x_k) = u(x_{k-1}) + m^{\$}(x_{k-1}, x_k) = \sum_{j=1}^{k-1} m^{\$}(x_{j,}\, x_{j+1})\\
&u(x_n) = u(x_{n-1}) + m^{\$}(x_{n-1}, x_n) = \sum_{j=1}^{n-1} m^{\$}(x_{j,}\, x_{j+1})
\end{aligned}$$

Since $m^{\$}(x_j, x_{j+1})$ is nonnegative, it follows that

$$u(x_1) \leq u(x_2) \leq u(x_3) \leq \ldots \leq u(x_n)$$

so that u preserves the order on S—that is, u preserves the relation R. Furthermore, we have

$$u(x_k) - u(x_{k-1}) = m^{\$}(x_{k-1}, x_k)$$

Thus,

$$(x_1, x_{x+1}) R^{\$} (x_j, x_{j+1})$$

if and only if

$$m^{\$}(x_1, x_{i+1}) \geq m^{\$}(x_j, x_{j+1})$$

if and only if

$$u(x_{i+1}) - u(x_i) \geq u(x_{j+1}) - u(x_j)$$

Hence, the measure u also preserves $R^{\$}$. This completes the proof of sufficiency. For necessity, see Exercise 29. ◇

To illustrate the procedure outlined in the proof, suppose we have a set of five elements with standard order x_1, x_2, x_3, x_4, x_5. Then the set T consists of four pairs,

$$T = \{(x_1, x_2), (x_2, x_3), (x_3, x_4), (x_4, x_5)\}$$

(see Fig. 7.2).

Suppose that examination of the relation $R^{\$}$ indicates that the standard ordering on T is

$$(x_3, x_4), \quad (x_2, x_3), \quad (x_4, x_5), \quad (x_1, x_2)$$

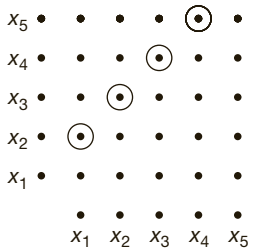


FIGURE 7.2 The heavy dots indicate the elements of $S \times S$. The members of T are circled.

so that $m^{\S}(x_3, x_4) \leq m^{\S}(x_2, x_3) \leq m^{\S}(x_4, x_5) \leq m^{\S}(x_1, x_2)$ for every measure $m^{\S}$ on the system $< T, R^{\S} >$. If scale values of 3, 5, 6, 7 are chosen for $m^{\S}$, then the proof of Theorem 5 defines a measure u by

$$u(x_1) = 0, \quad u(x_2) = 7, \quad u(x_3) = 12, \quad u(x_4) = 15, \quad u(x_5) = 21.$$

Now the measure u can be used to define a relation R^u on the full set $S \times S$. Define $(x, y)\, R^u (z, w)$ if and only if $u(y) - u(x) \geq u(w) - u(z)$. Note that $R^{\S}$ is a subset of R^u, so we might say that R^u *extends* $R^{\S}$. As an example, note that since

$$u(x_4) - u(x_2) = 15 - 17 = 8$$

while

$$u(x_5) - u(x_4) = 21 - 5 = 6$$

we have

$$(x_2, x_4)\, R^u\, (x_4, x_5)$$

The choice of scale values for $m^{\S}$ is, as we have seen earlier, unique only up to an order-preserving transformation. We might have chosen, with equal validity, scale values of 1, 2, 4, 8. With these values for $m^{\S}$, we obtain a measure v on S with

$$v(x_1) = 0, \quad v(x_2) = 8, \quad v(x_3) = 10, \quad v(x_4) = 11, \quad v(x_5) = 15.$$

As in the previous paragraph, we may use v to define a relation R^v on $S \times S$. Again, $R^{\S}$ will be a subset of R^v so that R^v also extends $R^{\S}$. Now, however, we will have $v(x_4) - v(x_2) = 11 - 8 = 3$, while $v(x_5) - v(x_4) = 15 - 11 = 4$ so that in this extension $(x_4, x_5)\, R^v\, (x_2, x_4)$.

One set of scale values for $m^{\S}$ is consistent with the student's judgment that there is a greater difference between x_2 and x_4 than between x_4 and x_5, while a different set of scale values is not. In this alternative approach, we have agreed not to ask the subject to make comparisons between pairs (x_2, x_4) and (x_4, x_5). This example shows that we cannot determine what judgment the student would make on these pairs solely on the information we have concerning the pairs in the set T. The Second Representation Theorem (Theorem 5) then gives a measure u that is a better reflection of the student's attitude toward the courses in the Winter Term than the ordinal measure m, but it does not completely answer the objections raised by the student complaint of Section IV.

VI. Some Historical Notes

Although it has been recognized since ancient times that measurement is essential to any scientific theory attempting to explain real-world phenomena, no attempt to study the foundations of measurement theory was made until the 20th century. The German mathematician Otto Ludwig Hölder (1859–1937) published an axiomatization for the measurement of mass in 1901. The general theory of measurement in physics was studied quite extensively by the British physicist Norman Robert Campbell (1880–1949). Campbell noted that the basic quantities measured by physicists all shared two common properties:

1. Given any two objects, it is always possible to decide which one "possesses" more of the quantity than the other.
2. There is an operation of combining any two objects that corresponds to the arithmetical operation of addition.

To cite one example, think of the process of determining lengths of a set of straight, rigid rods. If we place two rods side by side so that they coincide at one end, we determine which one is longer by examining the other end and observing which one extends farther. Thus, Property (1) is satisfied. For Property (2), note that two or more rods can be combined or *concatenated* by placing them end-to-end in a straight line. The concept of length dictates that the length of such a concatenated rod be the sum of the lengths of the component rods.

In the discussion of ordinal scales in this chapter, we saw how to axiomatize Property (1). Property (2) demands the extra condition that the measure of the concatenation of any two objects be equal to the sum of the measures of the subjects.

Campbell distinguished between two kinds of measurement, which he called "intensive" and "extensive." A measurement is *extensive* if the underlying quantity satisfies Properties (1) and (2) and *intensive* if it satisfies only Property (1). Most psychological and sociological attributes are intensive while most physical properties are extensive in nature.

Measurement theory is now an active branch of all the mathematical social sciences. Much of the current work in this area was stimulated by axiomatic studies undertaken by R. Duncan Luce and Patrick Suppes, beginning in the 1950s.

EXERCISES

II. What is Measurement?

Exercises 1–4 refer to the example of Section II.A.

1. Prove that $w(A) = w(B)$ if and only if A^*B and B^*A.
2. How would you describe, using the "*" notation, the fact that A is heavier than B?
3. Let A, B, and C be any three objects. Show that the following statements are all true:
   (a) A^*A.
   (b) If A^*B and B^*C, then A^*C.
   (c) Either A^*B or B^*A, or both.
4. If $w(A) = 1$ and two copies of object A exactly balance one copy of B in a pan balance, does it follow from the procedure outlined in II.A that $w(B) = 2w(A)$? Why? How would you modify the procedure to ensure this?
5. Determine which of the Examples 1–12 are reflexive, symmetric, transitive, and connected.
6. Let C be the set of all ordered pairs (a, b) of real numbers, Define $(a, b)\, R\, (c, d)$ if and only if $a > c$ and $b > d$. Is this relation transitive? Is it connected?

7. Find an example of a relation that is symmetric and transitive, but not reflexive. (Consider the relation "x is a sibling of y.")

8. Write xPy if (x, y) is an element of the relation P and $xP'y$ if (x, y) is not an element. A connected, transitive relation is sometimes referred to as a *weak order*. A *strong order* is a relation P that is transitive and satisfies an "asymmetry" condition: xPy implies $yP'x$. Show that any weak order is the union of two disjoint sets, one of which is a strong order and the other is an equivalence relation. An *equivalence relation* is a relation that is reflexive, symmetric, and transitive.

9. Is a strong order always connected? Can it be reflexive?

10. Consider the relations on the set of real numbers determined by the concepts of $>$, $\geq$, $<$, $\leq$, and $\neq$. Which are weak orders? strong orders? equivalence relations?

11. Let S be the set of integers. Define a relation R by xRy if and only if $x - y$ is a multiple of 5. Show that R is an equivalence relation.

12. Let S be the set of all adults in New England. Define a relation R by xRy if and only if x lives with y. Is R an equivalence relation?

13. Let R be an equivalence relation on a set S. Split S into subsets by agreeing to put x and y into the same set exactly when xRy. Show that this procedure partitions S into a collection of pairwise disjoint subsets. These subsets are called *equivalence classes*. Carry out this process with the relation of Exercise 11; how many equivalence classes are there?

14. Can you carry out the process of creating equivalence classes defined in Exercise 13 if R is not an equivalence relation? Why?

15. A *semi-order* is a relation P on a set S satisfying the following three axioms for all x, y, z, w in S:

(i) $xP'x$.

(ii) If xPy and zPw, then either xPw or zPy.

(iii) *If* xPy and zPw, then either xPw or wPz.

Prove that every semi-order is transitive.

16. Write out a proof for Theorem 1.

III. Simple Measures on Finite Sets

17. At what steps in the proof of Theorem 3 is the finiteness of the set S used?

18. Show that procedure of the proof of Theorem 3 leads to $m(z) = 0$, $m(y) = 1$, $m(x) = 2$ for the relation of Example 13.

19. For Example 13, show that the function with scale values $m(z) = -17$, $m(y) = \sqrt{23}$, $m(x) = 10$ also satisfies the definition of a measure.

20. Let S be the set $\{w, x, y, z\}$ and R the relation $\{(w, x), (x, y), (w, y), (w, z), (z, y), (x, z)\}$.

(a) Show that R is connected and transitive.

(b) Find a measure for this relational system.

21. Suppose that m and u are measures on the relational system $< S, R >$ where S is finite. Show that there is an order-preserving function f: $M \rightarrow M$ where $M = \{m(x_i)\}$ such that $u(x_i) = f(m(x_i))$ for all i.

22. How many different questions of the type "Do you prefer course i to course j?" must you ask a student to construct a preference ordering for a set of 70 courses?

23. Show that the relation of Example 9 is connected and transitive, but the system $< S, R >$ has no measure in the sense of Theorem 3.

24. A set S is said to be *countably infinite* if there is a one-to-one correspondence between the elements of S and the set of all positive integers. Show that Theorem 3 is true if S is a countably infinite set.

25. **(a)** Let S be the set of all ordered pairs (a, b) of real numbers such that $a = 0$ or 1 and $0 \leq b \leq 1$. Let R be the lexicographic order on S. Does the system $< S, R >$ have a measure?

(b) Let T be the set of all ordered pairs (a, b) of real numbers such that $a = 0$ and b is either between 0 and 1 or between 2 and 3. Let R be the lexicographic order on T. Does the system $< T, R >$ have a measure?

IV. Perception of Differences

26. Let u be a real-valued function defined on a set S. Define a relation R' on S by $xR'y$ if and only if $u(x) \geq u(y)$. Show that R' is necessarily reflexive, transitive, and connected. Will R' be symmetric?

27. Prove that condition (4) of Theorem 4 holds.

28. Verify the claims made about Example 14.

V. An Alternative Approach

29. Prove that the conditions in the statement of Theorem 5 are necessary.

30. Show that scale values of 3, 5, 6, 7 for $m^{\S}(x_3, x_4)$, $m^{\S}(x_3, x_4)$, $m^{\S}(x_2, x_3)$, $m^{\S}(x_4, x_5)$, $m^{\S}(x_1, x_2)$ yield the measure $u(x_1) = 0,\ u(x_2) = 7,\ u(x_3) = 12,\ u(x_4) = 15,\ u(x_5) = 21$.

31. Show that $R^{\S}$ is a subset of both R^u and R^v.

32. List the members of R^u and R^v explicitly.

33. Use scale values of 2, 3, 5, 10 instead of 3, 5, 6, 7 for the function $m^{\S}$ to determine a measure w. Show that in the extension R^w you have $(x_2, x_4)R^w(x_4, x_5)$ and $(x_4, x_5)R^w(x_2, x_4)$.

34. Concatenation on a set S may be defined formally as a function f from $S \times S$ to S. We denote $f(x, y)$ by $x\Delta y$.

(a) Give an example of a concatenation such that $x\Delta y \neq y\Delta x$.

(b) A concatenation is said to be *associative* if $x\Delta(y\Delta w) = (x\Delta y)\Delta w$ for all x, y, w in S. If Δ is an associative concatenation, show that the following definition is unambiguous: "If n is a positive integer, then nx is defined to be x concatenated with itself n times—e.g., $2x = x \Delta x$."

35. An *extensive measurement system* is axiomatically defined as a triple $< S, R, \Delta >$ where S is a set, R is a connected and transitive relation on S, and Δ is an associative concatenation on S satisfying the following four properties:

(i) If xRy, then $(x \Delta z)R(y \Delta z)$.

(ii) If $x \not R y$, then there exists a w in S such that $xR(y \Delta w)$ and $(y \Delta w)Rx$.

(iii) $(x \Delta y) \not R x$.

(iv) If xRy, then there is a positive integer n such that $yRnx$ for all x, y, z in S.

Prove the following Representation Theorem: If $< S, R, \Delta >$ is an extensive measurement system, then there is a real-valued function m defined on S such that

$$m(x) \leq m(y) \quad \text{if and only if} \quad xRy$$

and

$$m(x\Delta y) = m(x) + m(y) \quad \text{for all } x \text{ and } y \text{ in } S$$

36. Show that the function m guaranteed by the theorem in Exercise 35 is unique up to multiplication by a positive constant.

37. Show that the ordinary conceptions of length and weight satisfy the axioms of an extensive measurement system.

SUGGESTED PROJECTS

1. A *quasi-measure* on a relational system $< S, R >$ is a real-valued function m defined on S such that $m(x) \geq m(y)$ if xRy. We do not require that xRy whenever $m(x) \geq m(y)$. What are necessary and sufficient conditions for the existence of quasi-measures? What are the analogues of the Representation Theorems presented in this chapter for quasi-measures? Identify some real-world quasi-measures.

2. Find necessary and sufficient conditions on a relation defined on an infinite set that will guarantee the existence of a measure in the sense of Theorem 3. Keep in mind Exercises 23–25.

3. It has been argued that in many situations, observed equality relations may not be transitive. A person may judge rod x to be as long as rod y, which in turn is judged as long as rod z; yet x may be judged longer than z. Such judgments arise whenever the differences between x and y and between y and z are too small to be noticed. The combined difference, however, may be sufficiently large to make a difference between x and z noticeable. The classic example is a sequence of cups of coffee each containing one more grain of sugar than the previous cup. An observer could probably detect no difference in sweetness between two adjacent cups. If "equally sweet" is a transitive relation, then we would have to conclude that a cup with no sugar in it is as sweet as one in which 10 teaspoons of sugar have been dissolved!

To handle such situations, R. Duncan Luce introduced the idea of a semi-order as the type of relation to capture the notion of strict preference (see Exercise 15). If P is a semi-order, then an indifference relation I can be defined by xIy if and only if neither xPy nor yPx. Show that I is reflexive and symmetric, but not necessarily transitive. Prove the following Representation Theorem: If P is a semi-order on a finite set S, then there is a real-valued function f

defined on S and a positive number δ such that for all x and y in S,

$$f(x) > g(y) + \delta \quad \text{if and only if} \quad xPy.$$

The constant δ may be interpreted as a single "just noticeable difference" unit. Is this Representation Theorem true for infinite sets?

4. Some mathematical psychologists have investigated attributes that appear to have a property somewhat analogous to a physical concatenation operation. This is the property that for each pair of objects x and y, there is a third object that lies "halfway" between x and y in terms of possession of the attribute under study. For example, a subject may be presented with two tones of different loudness and asked to adjust a variable tone until its subjective intensity "bisects" the loudness of the given pair.

A *bisection system* is a triple $< S, R, B >$ where R is a connected, transitive relation on a set S, and B, the bisection operation, is a function from $S \times S$ to S. The element $B(x, y)$ is interpreted as the subject "midpoint" between x and y.

Find a reasonable set of axioms on the function B that guarantees the existence of a real-valued scale f defined on S that preserves the relation R and such that the scale value assigned to the "midpoint" is a weighted average of the scale values of the "endpoints."

You can find a listing of references and suggestions for additional reading on the book's website, www.wiley.com/college/olinick

Introduction to Utility Theory

CHAPTER 8

Some reckon time by stars,
And some by hours;
Some measure days by Dreams,
And some by flowers;
My heart alone records
My days and hours.

—Madison Cawein

I. Introduction

This chapter continues the axiomatic discussion, begun in Chapter 7, of certain aspects of measurement theory. We consider again the problem that motivated the development of the material in the preceding chapter from a new point of view. The problem is to construct a numerical measurement of "happiness"; in particular, to assign numbers that measure how happy a particular student would be if she were assigned various different courses by the college's registrar.

The point of view of this chapter is called *utility theory*. The theory dates back at least 200 years to a time when nobles of the French court asked mathematicians for advice on how to gamble. Quite a rich theory has been developed, and various aspects of it have been tested experimentally in situations requiring decision making with incomplete knowledge.

Consider the set S of possible choices of courses to which the student might be assigned. Using the mechanisms of Chapter 7, or some other scheme, it is determined that the student prefers course x over course y and course y over course z. Utility theory aims to assign numerical weights to these preferences.

Suppose we offer the student a choice: she may have course y, her intermediate choice, or she may flip a coin. If the coin comes up heads, she gets course x, while if it comes up tails, she gets course z. Which option does she prefer: the certainty of y or the gamble between x and z?

If the coin is weighted so that it always comes up heads, then she will certainly always prefer the gamble: there is a certainty that she will receive her first choice. If the coin is weighted so that it always lands with tails showing, then she will forego the gamble and take course y.

Suppose the coin is an honest one, so that the likelihood of winning x on the flip is the same as winning z. What can we say if the student prefers course y to a gamble with an

honest coin? We should be able to deduce that she perceives the difference in happiness between x and y to be less than the difference between y and z. (Why?)

The theory of utility is based on the assumption that there is a way of weighting the coin so that the student has no preference between the gamble and the certainty. This ideal weight can then be translated into a measure of happiness about the course y.

The remainder of this chapter develops a mathematical model reflecting the ideas of this previous paragraph.

II. Gambles

In developing utility theory, it is convenient to introduce two binary relations on the set S, one based on strict preference (P) and one on indifference (I).

DEFINITION If S is a set and P is a binary relation on S, define the *indifference relation* xIy for any pair of elements x and y of S if and only if neither xPy nor yPx.

The first theorem indicates why this is an important relation.

THEOREM 1 Suppose u is a real-valued function defined on S such that $u(x) > u(y)$ if and only if xPy. Then the following conditions hold:

1. Given any two elements x and y of S, exactly one of three possibilities is true: xPy yPx, or xIy.
2. P is transitive.
3. I is reflexive, symmetric, and transitive.
4. If xPy and yIz, then xPz.
5. If xIy and yPz, then xPz.

Proof of Theorem 1 The proof depends upon the elementary-order properties of the real numbers and is similar in spirit to the proofs studied in Chapter 7. We prove condition (4) here; we leave the other properties as an exercise for the reader.

Suppose then that xPy and yIz. Consider the numbers $u(y)$ and $u(z)$. If $u(y) > u(z)$, then we would have yPz, while if $u(z) > u(y)$, we must have zPy. Since neither yPz nor zPy, we must have $u(y) = u(z)$. But xPy gives $u(x) > u(y)$ and hence $u(x) > u(z)$. Thus, xPz. ◇

We come now to the crucial definition for utility theory.

DEFINITION Let x and y be any two elements of a set S and let p be a number, $0 \le p \le 1$. Then the symbol $px + (1 - p)y$ represents the *gamble*, or *lottery*, that has two possible outcomes, x and y, with probabilities p and $1 - p$, respectively.

The phrase "probability" p may be interpreted as meaning roughly that if the gamble is repeated a very large number of times, we may expect outcome x to occur about $100p$ percent of the time. For example, think of the symbol $.25x + .75y$ as representing the gamble of flipping a coin that has been weighted so that it turns up heads (outcome x), on the

average, 25% of the time. Alternatively, imagine a coin that comes up tails three times as often as heads, but has no predictable pattern.

The gamble $1x + 0y$ is simply denoted as x. Gambles with three or more possible outcomes may also be defined. The symbol $px + qy + (1 - p - q)z$ would represent a gamble with three possible outcomes, x, y, and z, having probabilities p, q, $1 - p - q$, respectively, where p and q are nonnegative numbers whose sum is at most 1.

It is also possible to consider gambles in which one of the possible outcomes is itself a gamble. Suppose, for example, you are offered the following proposition.

Flip a coin. If it comes up heads, you receive a new automobile (outcome x). If it comes up tails, then you roll a die. If the die shows a "3," you win a radio (outcome y); otherwise you lose \$10,000 (outcome z).

Assuming that the coin and the die are "honest," this compound gamble can be represented as

$$\frac{1}{2}x + \frac{1}{2}\left(\frac{1}{6}y + \frac{5}{6}z\right)$$

If this particular gamble were repeated a large number of times, say 12,000, what should happen? The coin should turn up heads about 6,000 times and tails 6,000 times. For the 6,000 rolls of the die, we should see a "3" about 1,000 times and one of the other five numbers about 5,000 times. Thus, we should expect x to be the outcome about $\frac{1}{2}$ of the time, y about $\frac{1}{12}$ of the time, and z about $\frac{5}{12}$ of the time. This means that the gamble should be equivalent to a gamble with three outcomes x, y, z, having respective probabilities of $\frac{1}{2}$, $\frac{1}{12}$, and $\frac{5}{12}$ —that is, the gamble

$$\frac{1}{2}x + \frac{1}{2}\left(\frac{1}{6}y + \frac{5}{6}z\right)$$

is equivalent to the gamble

$$\frac{1}{2}x + \frac{1}{12}y + \frac{5}{12}z$$

Since the two gambles are equivalent, any reasonable person should be indifferent if offered a choice between them. There is no reason to prefer one of the gambles to the other.

III. Axioms of Utility Theory

A utility measure on a set S is determined by establishing preferences among the elements of the set of all gambles with outcomes in S. Some of these preferences will necessarily be dictated by the preference and indifference relations, P and I, which hold among the elements of S. For example, if the student prefers course x to y, then she should prefer the gamble $.7x + .3y$ to the gamble $.7y + .3x$, since the preferred outcome is more likely in the first gamble than in the second.

Utility theory assumes that there are binary relations P and I on the set of gambles with outcomes in S that are consistent with the already established preference and

indifference relations on S, satisfy the conditions (1)−(5) of Theorem 1, and also satisfy some additional reasonable axioms.

The first three axioms simply assert that the student should be indifferent if offered a choice between essentially equivalent gambles. Formally, these axioms look like the following.

For all x, y, z in S and all real numbers p, $0 \le p \le 1$,

AXIOM 1 $[px + (1-p)y]I[(1-p)y + px]$.

AXIOM 2 $[px + (1-p)\{qy + (1-q)z\}]I[px + (1-p)qy + (1-p)(1-q)z]$ where q is any number $0 \le q \le 1$.

AXIOM 3 $[px + (1-p)x]Ix$.

Consider now two gambles: $.3x + .7z$ and $.3y + .7z$. In both gambles, outcome z has probability $\frac{7}{10}$ and the other outcome has probability $\frac{3}{10}$. Which gamble would the student prefer? Clearly it should depend on her preference between outcomes x and y. If she prefers x to y, then she should prefer the first gamble to the second and if she is indifferent between x and y, then there is no reason for her to prefer one gamble over the other. This example indicates that two additional axioms are reasonable.

AXIOM 4 If xPy, then for any $p > 0$, $[px + (1-p)z]P[py + (1-p)z]$.

AXIOM 5 If xIy, then for any p, $[px + (1-p)z]I[py + (1-p)z]$.

To see how these axioms fit together, suppose that the student prefers course x to course z. Then if she is offered two different gambles with x and z as the outcomes, she should prefer the gamble in which there is a greater likelihood of outcome x. Axioms 1−5 enable us to prove this result.

THEOREM 2 If xPy and p and q are numbers with $0 < q < p < 1$, then $[px + (1-p)y]\ P[qx + (1-q)y]$.

Proof of Theorem 2 Since $0 < q < p < 1$, we have $0 < p - q < 1 - q$, and by Axiom 3,

$$yI\left[\frac{p-q}{1-q}y + \frac{1-p}{1-q}y\right]$$

Axiom 4 gives

$$\left[\frac{p-q}{1-q}x + \frac{1-p}{1-q}y\right]P\left[\frac{p-q}{1-q}y + \frac{1-p}{1-q}y\right]$$

Let z denote the gamble

$$z = \frac{p-q}{1-q}x + \frac{1-p}{1-q}y$$

so that we have, using condition (4) of Theorem 1, zPy. The gamble $px + (1 - p)y$ is equivalent to the gamble $qx + (1 - q)z$ so that

$$[px + (1 - p)y]\ I[qx + (1 - q)z]$$

Using Axiom 4 again, we have $[qx + (1 - q)z]\ P[qx + (1 - q)y]$. Transitivity of P completes the proof. ◇

We will define a utility measure on the set of all gambles with outcomes in S in such a way that the measure of one gamble will be greater than the measure of another exactly when the first gamble is preferred to the second. The measures will be identical if there is indifference between the gambles.

To establish the existence of such a measure, we require one additional axiom. Consider the gamble $px + (1 - p)z$. If $p = 0$, then this gamble is equivalent to the certainty of outcome z, while if $p = 1$, then the gamble is identified with the outcome x. It seems reasonable that a slight change in the value of p should result only in a small change in the utility measure of the gamble. Thus, as p varies continuously from 0 to 1, the utility measure of $px + (1 - p)z$ should vary continuously from the measure of z to the measure of x. If y is some outcome whose measure lies between the measure of z and the measure of x, then the intermediate value theorem of elementary calculus would assert that there is at least one value of p for which the measure of y is equal to the measure of $px + (1 - p)z$. (See Fig. 8.1.)

The final axiom captures this idea in terms of the preference and indifference relations:

AXIOM 6 If x, y and z are elements of S with xPy and yPz, there is at least one number p, $0 \le p \le 1$, such that $[px + (1 - p)z]Iy$.

Axiom 6 asserts that there is always some gamble that is indifferent to the certainty of y and whose prescribed outcomes are two events, one preferred to y and one preferred less than y. In fact, there is exactly one such gamble.

THEOREM 3 If xPy, yPz and $[px + (1 - p)z]Iy$, then p is unique. Furthermore, p is strictly between 0 and 1—that is, $0 < p < 1$.

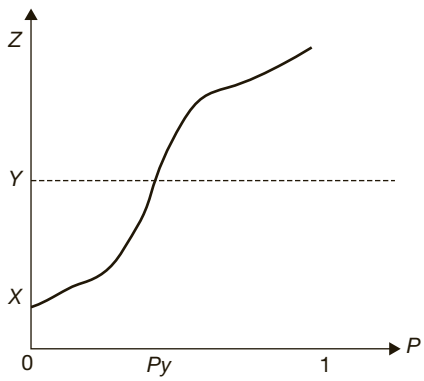


FIGURE 8.1 Graph of the measure of the gamble $px + (1 - p)z$ as a function of p. The letters X, Y, Z indicate the measures of the gambles x, y, z, respectively. If the measure is a continuous function of p and if Y lies between X and Z, then for some p_y between 0 and 1, the measure of $p_y x + (1 - p_y)z$ should be Y.

Proof of Theorem 3 If $p = 0$, then the gamble $px + (1 - p)z$ is equivalent to the outcome z and we would have zIy, which violates the assumption that yPz. A similar argument shows that p cannot be equal to 1.

Now let q be any number between 0 and 1 that is not equal to p, but such that $[qx + (1 - q)z]Iy$. The transitivity of I implies that

$$[qx + (1 - q)z]I[px + (1 - p)z]$$

This indifference, however, contradicts Theorem 2. ◇

IV. Existence and Uniqueness of Utility

A. Existence

The axioms and theorems of Section III provide the ammunition to state and prove a representation theorem for utility functions. Compare Theorem 4 below with Theorems 3 and 5 of Chapter 7.

THEOREM 4 (THIRD REPRESENTATION THEOREM) There is a real-valued function u defined on S such that

1. $u(x) > u(y)$ if and only if xPy, and
2. $u(px + (1 - p)y) = pu(x) + (1 - p)u(y)$ for every pair of elements x and y of S and every real number p, $0 \leq p \leq 1$.

Proof of Theorem 4 Suppose first that xIy for all x and y in S. In this case, let $u(x) = 0$ for all elements x of S. It is a triviality that (1) and (2) are satisfied. If there is not complete indifference, then find some pair of elements x_1 and x_0 in S with $x_1 \; P \; x_0$. Define $u(x_0) = 0$ and $u(x_1) = 1$. Now let x be any other element in S. There are five possibilities to consider:

a. xIx_0

b. xIx_1

c. x_1Px and xPx_0

d. xPx_1

e. x_0Px

We show how to define $u(x)$ in each of these cases. See Fig. 8.2. ◇

In this manner, we define $u(x)$ for each element x of S. It must be shown that this function satisfies conditions (1) and (2) in the statement of the theorem. Toward this end, let x and y be any two elements in S. There are actually 25 cases to consider, depending on whether each of the two elements, x and y, lies in cases (a), (b), (c), (d), or (e). In some of the cases, it is immediate that conditions (1) and (2) must hold. This happens, for example, if

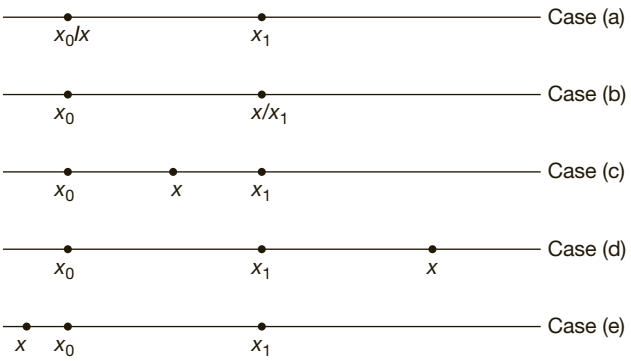


FIGURE 8.2 Schematic representation of the five cases in the Third Representation Theorem.

Case (a) If xIx_0, let $u(x) = u(x_0) = 0$.

Case (b) If xIx_1, let $u(x) = u(x_1) = 1$.

Case (c) If x_1Px and xPx_0, then there is a unique number $p_0 < p < 1$, such that $[px_1 + (1-p)x_0]Ix$. Define $u(x) = p$. Note that $u(x_0) < u(x) < u(x_1)$.

Case (d) If xPx_1, then x_1 is intermediate between x_0 and x. There is a unique number q, $0 < q < 1$, such that $[qx + (1-q)x_0]x_1$. Define $u(x) = 1/q$. Note that $u(x) > 1 = u(x_1)'$.

Case (e) If x_0Px, then x_0 is intermediate between x and x_1 and there is a unique number r, $0 < r < 1$, such that $[rx_1 + (1-r)x]Ix_0$. Define $u(x) = r/(r-1)$. Note that $u(x) < 0 = u(x_0)$ in this case.

both elements are indifferent to x_0. We will consider in detail a typical nontrivial case: the one in which both x and y belong to Case (c).

Suppose, then, that $u(x) = p_1$ and $u(y) = p_2$. Let A be the gamble $p_1x_1 + (1-p_1)x_0$ and let B be the gamble $p_2x_1 + (1-p_2)x_0$. By definition of the function u, we have AIx and BIy.

If $p_1 = p_2$, A and B are the same gamble, so that AIx and AIy. Since I is a symmetric and transitive relation, we have xIy.

If $p_1 > p_2$, then Theorem 2 (with $p = p_1$, $q = p_2$, $x = x_1$, $y = x_0$) gives APB. Axioms 4 and 5 then imply that xPy.

Similarly, should p_1 be less than p_2, the application of Theorem 2 and Axioms 4 and 5 yields yPx.

This establishes condition (1).

To prove that condition (2) is true, let p be any real number between 0 and 1. Strictly speaking, we have not yet defined the utility measure of the gamble $px + (1-p)y$. On the other hand, since x_1PxPx_0 and x_1PyPx_0, Axiom 4 gives $x_1P\ [px + (1-p)y]Px_0$. The same definition of u as above can be used if there is a number p^* such that $[px + (1-p)y]I[p^*x_1 + (1-p^*)x_0]$. This is easy to find.

Since AIx and BIy, Axiom 5 yields $[px + (1-p)y]I[pA + (1-p)B]$ but

$$\begin{aligned}[pA + (1-p)B]I[p\{p_1Ix_1 + (1-p)x_0\} + (1-p)\{p_2x_1 + (1-p_2)x_0\}] \\ I\{pp_1 + (1-p)p_2\}x_1 + \{p(1-p_1) + (1-p)(1-p_2)\}x_0 \\ I\{pp_1 + (1-p)p_2\}x_1 + \{I - (pp_1 + (1-p)p_2)\}x_0\end{aligned}$$

so that $p^* = pp_1 + (1-p)p_2$ and $u(px + (1-p)y) = p^* = pu(x) + (1-p)u(y)$.

The proofs that conditions (1) and (2) hold in the remaining cases are quite similar and are left to the reader.

B. Uniqueness of Utility

How much freedom is there in the choice of scale values for the utility function? Recall that the scale value for a measure on a binary relational system $< S, R >$ was unique up to an order-preserving transformation. For a utility function, there is less freedom. The scale is unique up to changes by a positive linear transformation.

DEFINITION A real-valued function L defined on a set T of real numbers is a *positive linear transformation* if there are constants α and β where $\alpha > 0$ such that $L(t) = \alpha t + \beta$ for every element t in T.

As an example, let T be the set of numbers between 0 and 100 and consider the positive linear transformation $L(t) = (9/5)t + 32$. Then $L(0) = 32$ and $L(100) = 212$. This linear transformation may be familiar to you as the one that converts Celsius temperatures to Fahrenheit temperatures.

THEOREM 5 If u is a utility function in the sense of Theorem 4 and L is a positive linear transformation, then the composition $u^* = L \circ u$ is also a utility function.

Proof of Theorem 5 Suppose $L(t) = \alpha t + \beta$ so that $u^*(x) = L(u(x)) = \alpha u(x) + \beta$. Then the inequality

$$u^*(x) > u^*(y)$$

is the same as

$$\alpha\, u(x) + \beta > \alpha\, u(y) + \beta$$

or

$$\alpha\, u(x) > \alpha\, u(y)$$

Since α is positive, we have $u^*(x) > u^*(y)$ exactly when $u(x) > u(y)$—that is, exactly when xPy.

Similarly,

$$\begin{aligned} pu^*(x) + (1-p)u^*(y) &= p(\alpha u(x) + \beta) + (1-p)(\alpha u(y) + \beta) \\ &= \alpha\, pu(x) + \alpha(1-p)u(y) + (p + 1 - p)\beta \\ &= \alpha[pu(x) + (1-p)u(y)] + \beta \\ &= \alpha[u(px + (1-p)y)] + \beta \\ &= u^*(px + (1-p)y) \end{aligned}$$

the next to the last equality holding since u satisfies condition (2) of Theorem 4. Thus, the function u^* also satisfies condition (2). This completes the proof that u^* is a utility function. ◇

THEOREM 6 If u and v are utility functions in the sense of Theorem 4, then there is positive linear transformation L so that $v = L \circ u$.

Proof of Theorem 6 Since $x_1 \; P \; x_0$, we must have $v(x_1) > v(x_0)$ so that $v(x_1) - v(x_0) > 0$. Define the positive linear transformation $L(t) = \alpha t + \beta$ by $\beta = v(x_0)$ and $\alpha = v(x_1) - v(x_0)$.

Now let x be any element of S. There are five cases to be considered, depending on which of the possibilities (a)−(e) of Theorem 4 is true. We give the proof in two cases:

Case (*c*) Here $x_1 P x P x_0$.

If $u(x) = p$, then $xI[px_1 + (1-p)x_0]$ so that $v(x) = v(px_1 + (1-p)x_0)$ since v preserves I. Now the right-hand side can be written as

$$pv(x_1) + (1-p)v(x_0)$$

since v satisfies condition (2) of Theorem 4. Thus,

$$\begin{aligned} v(x) &= pv(x_1) + (1-p)v(x_0) \\ &= p(\alpha+\beta) + (1-p)\beta = p\alpha + \beta = \alpha u(x) + \beta = L(u(x)) \end{aligned}$$

Case (*d*) Here $x P x_1$.

If $u(x) = p$, set $q = 1/p$ so that $x_1 I[qx + (1-q)x_0]$. Again we have $v(x_1) = v(qx + (1-q)x_0) = qv(x) + (1-q)v(x_0)$ since v preserves I and satisfies condition (2) of a utility function. Solving for $v(x)$, we obtain

$$\begin{aligned} v(x) &= (1/q)[v(x_1) - (1-q)v(x_0)] \\ &= p[\alpha + \beta - (1-q)\beta] \\ &= p(\alpha + q\beta) = p\alpha + (pq)\beta \\ &= p\alpha + \beta (\text{since } pq = 1) \\ &= \alpha u(x) + \beta = L(u(x)) \end{aligned}$$

The remaining three cases are left for the reader to verify. ◇

The proof of Theorem 6 indicates that once values $v(x_1)$ and $v(x_0)$ are chosen for any two nonindifferent outcomes x_1 and x_0, the utility function v is completely determined, not only for the elements of S, but for all gambles with outcomes in S.

V. Classification of Scales

Scales that are invariant under positive linear transformations are called *interval scales*. Recall that a scale that is invariant under monotone transformations, such as that given in the First Representation Theorem of Chapter 7, was called an ordinal scale. The idea of classifying scales and measurement functions by the types of transformations that preserve the underlying binary relations is due to the American psychologist Stanley

Smith Stevens (1906–1973). In his papers, dating from 1946, Stevens isolates five major types of scales: nominal, ordinal, interval, ratio, and absolute. We list these in order of increasing restrictions on the type of transformation permitted.

Nominal scales are invariant under all one-to-one transformations and are used when the basic empirical observation to be measured is the determination of equality; a standard example is the assignment of numbers to the jerseys of football players on a team: two different players wear two different numbers. As we have seen, we employ the *ordinal scale* when the empirical operation is the determination of "greater or less."

Interval scales reflect the operation of determining ratios of differences. The measurement of temperature is a good example of an interval scale measurement. The Fahrenheit and Celsius scales are positive linear transformations of each other. As we have seen, the transformation $L(t) = (9/5)t + 32$ converts Celsius to Fahrenheit. Conversely, the transformation $L^*(t) = (5/9)t - (5/9)\,32$ converts Fahrenheit to Celsius.

Ratio scales are those invariant under similarity transformations—that is, functions of the form $S(t) = \alpha t$ where α is a positive constant. The underlying empirical observation here is the determination of ratios and is exemplified by the measurement of length, weight, density, loudness, and pitch. Thus, it makes sense to say that one rod is twice as long as another, while it does not make sense to assert that one body of water is twice as hot as another.

In the *absolute scale*, only the trivial identity transformation is permitted. Counting, interpreted as an act of measurement, is an example of an absolute scale.

Of the three Representation Theorems, the final one, involving the idea of a utility function and an interval scale, gives the most information about how a subject assesses a set of objects for the degree to which a certain attribute is present. To find out whether the student prefers course x to course y, it is only necessary to compare the numbers $u(x)$ and $u(y)$. Furthermore, the utility function predicts how the student would rate the differences between pairs of courses.

To see how this is done, suppose the student's ordering of courses is x, y, z, w so that $u(x) < u(y) < u(z) < u(w)$. Offer the student two gambles:

$$.5x + .5w \text{ and } .5y + .5z.$$

If the student prefers the first gamble to the second, then

$$u(.5x + .5w) > u(.5y + .5z)$$

so that

$$.5u(x) + .5u(w) > .5u(y) + .5u(z)$$

or

$$u(x) + u(w) > u(y) + u(z)$$

implying that

$$u(w) - u(z) > u(y) - u(x)$$

Thus, the student perceives the difference between z and w to be greater than the difference between x and y. Should the student prefer the second gamble to the first, we can make the opposite conclusion.

VI. Interpersonal Comparison of Utility

Suppose the approach of utility theory is used to formulate the Registrar's Problem of Chapter 7. Determine for each student numerical scale values measuring that student's satisfaction with the courses being offered. One further refinement is necessary before we try to solve the Registrar's Problem.

Examine a very simple example. Suppose every student except Bob and Fred has been assigned to some course. There are two enrollment slots left, one in a course on Russian Literature and one in a class called Presidential Campaigning. Utility values for these students and courses are given in Table 8.1.

Which student should be assigned to which course? There are only two possibilities available. If Bob is assigned to Russian Literature and Fred to Presidential Campaigning, then adding the corresponding utilities gives $50 + 1.7 = 51.7$, whereas if Fred is enrolled in the literature course and Bob in the other one, we have $1.6 + 70 = 71.6$.

It seems that the second assignment increases the total satisfaction of the student body more than the first. Recall, however, that the scale values for a utility function are unique only up to a linear transformation. If each of Fred's scale values is multiplied by 300, say, his preferences are still preserved. The resulting scale values are indicated in Table 8.2.

With these values, the first assignment (Bob in Russian Literature, Fred in Presidential Campaigning) increases satisfaction by $50 + 510 = 560$, while the alternative assignment increases total happiness by only $70 + 480 = 550$. Now the first assignment seems better.

This example shows that the particular choices of scale values affect in a crucial way the solution of the Registrar's Problem. There is no clear, unambiguous solution to the problem, because there is as yet no single absolute scale against which to measure the utilities of different individuals. Difficulties arise because the given information does not indicate whether Bob's rating a course with 70 is a particularly high or particularly low rating for him. A similar comment holds for Fred's ratings. It makes a great difference in the assignment of courses if Bob's highest rating for a course is a 700 or if it is only a 75. There are, however, several ways of attempting to avoid the indicated difficulty. One way is to construct an absolute scale that forces the interpersonal comparison of utilities.

Table 8.1

	Bob	Fred
Russian Literature	50	1.6
Presidential Campaigning	70	1.7

Table 8.2

	Bob	Fred
Russian Literature	50	480
Presidential Campaigning	70	510

Suppose that the lowest rating a student gives to any course is A and the highest rating is B. Then the linear transformation

$$L(t) = \frac{1}{B-A}t - \frac{A}{B-A}$$

has the property that $L(A) = 0$, $L(B) = 1$, and $0 \le L(t) \le 1$ for all t, $A \le t \le B$. Thus, if a student's utility function is bounded above and below by numbers B and A, respectively, it is possible to rescale his utility measure so that all the scale values lie between 0 and 1. A measure of 70 will be rescaled closer to 1 if the highest rating is a 75 than it will be if the highest rating is a 700.

Continuing with the example, suppose all Bob's original scale values lie between 0 and 100. Then $A = 0$, $B = 100$, and the required positive linear transformation is $L_B(t) = t/100$. If Fred's original choice of scale values ran between -1 and 2, then the normalizing transformation is

$$L_F(t) = \frac{1}{2-(-1)}t - \frac{-1}{2-(-1)} = \frac{t+1}{3}$$

These transformations give $L_B(50) = .5$, $L_B(70) = .7$, $L_F(1.6) = .866$, and $L_F(1.7) = .9$. The normalized scale values are also indicated in Table 8.3.

With these values, the first assignment results in an increase of satisfaction of .5 + .9 = 1.4 while the second assignment gives .7 + .866 = 1.566. The second assignment is preferred to the first if all scales are normalized to lie between 0 and 1. In formulating the Registrar's Problem, we will always choose the numbers r_{ij} to be such normalized utility measurements.

This normalization—which makes possible an interpersonal comparison of utilities—is possible whenever all individuals in the group being studied have original bounded utility functions. Will utility measures necessarily be bounded? The Third Representation Theorem makes no restriction on the size of the set S. It may be finite or infinite. If the domain of a utility function is a finite set S (this occurs quite frequently in applications such as the Registrar's Problem), then the scale values will certainly be bounded. Even in the case that S is finite, however, the set of all gambles with outcomes in S will be infinite. The utility measure has as its domain the infinite set of gambles. There are also many instances in which it is necessary or convenient to assume that S is infinite. In such cases, there are no *mathematical* grounds for concluding that the utility function will necessarily be bounded. Sometimes there are empirical considerations for concluding that utility must be bounded.

Consider, for example, a game with an infinite number of possible outcomes, some desirable and others not. A player in the game wants to construct a utility scale that measures the value to him of each of the outcomes. In his doctoral dissertation at Princeton University, John R. Isbell developed a theory of cooperative games that is predicated on the assumption that such a utility scale will always be bounded. "Introspection convinces the

Table 8.3

	Bob	Fred
Russian Literature	.5	.866
Presidential Campaigning	.7	.9

present author firmly," he wrote, "that there is no prospect so desirable as to be worth an even bet of his life. He who claims a utility space unbounded above must in principle stand ready to bet his life at any odds provided the price is right."

To illustrate Isbell's argument more precisely, suppose the utility scale is unbounded above. This means that given any number M, then there exists some possible outcome z with $u(z) > M$. Consider, then, the two events

x: you win \$10,000

and

y: your head is chopped off.

Each of these outcomes is assigned a utility, $u(x)$ and $u(y)$. Let M be the number $10u(x) - 9u(y)$ and let z be an outcome with $u(z) > M$.

Since $u(z) > 10u(x) - 9u(y)$, we have $u(z) + 9u(y) > 10u(x)$ or $.1u(z) + .9u(y) > u(x)$. This last inequality implies that the gamble $.1z + .9y$ is preferred to the outcome x. The conclusion: there is some outcome so good that you are willing to give up a sure chance of winning \$10,000 to take a gamble of winning that outcome when there is a 90 percent chance that you will lose the gamble and, with it, your head. Isbell would argue that such an outcome is inconceivable.

One can construct a similar argument that "proves" that the utility scale is also bounded below. It should be noted that not all utility theorists accept the validity of such arguments.

VII. Historical and Biographical Notes

A. Utility Theory

Utility theory traces its ancestry back to the efforts of economists and mathematicians to develop an applicable theory of how a rational person ought to behave in the face of uncertainty and how, in fact, such a person does act. It was thought for a time that in economic situations people would act to maximize the expected value of money that would accrue to them. Thus, the gamble of winning \$10 if a fair coin lands heads and winning nothing if it lands tails shows an expected value of

$$\left(\frac{1}{2}\right)(\$10) + \left(\frac{1}{2}\right)(\$0) = \$5.$$

The rational man, under such a theory, should behave toward this gamble as if it were worth \$5.

It eventually became apparent that there are many instances when this idea is not applicable. Daniel Bernoulli (1700−1782), a member of the illustrious Swiss family that produced eight mathematicians in three generations, presents one: "Let us suppose a pauper happens to acquire a lottery ticket by which he may with equal probability win either nothing or 20,000 ducats. Will he have to evaluate the worth of the ticket as 10,000 ducats; and would he be acting foolishly, if he sold it for 9,000 ducats?"

In a paper written in 1790, Bernoulli explored the idea that the utility of money—not its actual value—is what people attempt to maximize. He argued that the utility of a fixed amount of money was different for a pauper than for a rich man. A single dollar is more

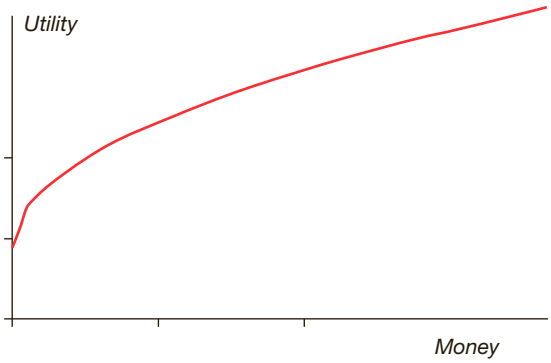


FIGURE 8.3 A possible graph of utility of money as a function of amount of money. Each increase in money increases utility, so the function is monotonically increasing. Fixed increases in money bring smaller increases in utility as money increases. Thus, the rate of change of utility is negative and the graph of function must be concave down.

precious to the poor man than to the millionaire; the poor man would feel the loss of a dollar more than the rich man. The difference in the utilities of \$10 and \$11 is greater, Bernoulli believed, than the difference in the utilities of \$1,000 and \$1,001. In general, a fixed increase in cash results in an ever smaller increase in utility as the basic cash wealth to which the increase is applied is made larger. In mathematical terms, this says that the graph of utility as a function of money is concave. See Fig. 8.3.

B. John Von Neumann and Oskar Morgenstern

"In John von Neumann's death on February 8, 1957, the world of mathematics lost a most original, penetrating, and versatile mind. Science suffered the loss of a universal intellect and a unique interpreter of mathematics, who could bring the latest (and develop latent) applications of its methods to bear on problems of physics, astronomy, biology, and the new technology."

Public domain

John Von Neumann

Public domain

Oskar Morgenstern

These are the words of the noted mathematician Stanislaw M. Ulam, and they express a judgment of Von Neumann universally shared by mathematicians and scientists who know his work. Von Neumann made important contributions to quantum physics, meteorology, the development of the atomic bomb, the theory and applications of high-speed computing machines, and to economics through the theory of games of strategy. He also made a number of outstanding discoveries in pure mathematics in the areas of measure and ergodic theory, continuous geometry, operator theory, topological groups, logic, and set theory.

Von Neumann was born in Budapest, Hungary, on December 28, 1903, and was the eldest son of a prosperous banker. His brilliant mind was revealed early. It is said that at the age of 6 he could divide two eight-digit numbers in his head, by age 8 he had mastered calculus, and by age 12 he was reading and understanding advanced books on function theory. His first published paper was written when he was 17, and the definition of ordinal number he created at age 20 is the one that is now universally used by mathematicians.

During the years 1922−1926, Von Neumann was registered as a student of mathematics at the University of Budapest but spent most of his time at the Eidgenossiche Technische Hochshule in Zurich, studying chemistry at the urging of his family who were doubtful of the financial future of a mathematician. Von Neumann was awarded his doctorate in mathematics from Budapest at about the same time he received his diploma in chemical engineering in Zurich.

Von Neumann first came to the United States in 1930 as a visiting lecturer at Princeton University. He became a professor there the next year and served on the faculty until 1933. In that year, the famous Institute for Advanced Study was founded. Among the first six professors given lifetime appointments in the Institute's School of Mathematics were Albert Einstein and John Von Neumann.

For most of the rest of the 1930s, Von Neumann's work was concentrated on pure mathematics. In 1940, however, there was a sharp break in his scientific work. As Paul Halmos describes it, "Until then he was a topflight pure mathematician who understood physics; after that he was an applied mathematician who remembered his pure work."

From 1940 until his death, Von Neumann served as a consultant to the Los Alamos Scientific Laboratory, the Naval Ordinance Laboratory, the Oak Ridge National Laboratory, and other military and civilian agencies. He was appointed to the Scientific Advisory Board of the Air Force and served as a member of the U.S. Atomic Energy Commission.

The technological development of the last generation that has had the greatest impact on society has been the high-speed electronic computer. Here, too, John Von Neumann played a critical role as a pioneer. He formulated the methods of translating a set of procedures into a language of instructions for a computing machine, made important contributions to the engineering of the first computers, and analyzed the question of whether machines could successfully imitate randomness or become self-reproducing automata.

Game theory as a model for the study of cooperation and competition has as its foundation a paper of Von Neumann written in 1928. His interest in this area was rekindled when the Austrian economist Oskar Morgenstern came to Princeton. Their intensive collaboration in the early years of World War II produced the 600-page *Theory of Games and Economic Behavior*. See Chapter 16 for a fuller treat of game theory.

Prior to Von Neumann and Morgenstern, mathematical economics had relied heavily on the techniques of mathematical physics and a rather shaky analogy between mechanics and economics. Von Neumann's innovations were to introduce the mathematical tools of axiomatization, convexity and combinatorics, and the fresh vie*wp*oint of analyzing

economic problems as games of strategy. The book also sparked new research into utility theory by mathematicians, economists, and psychologists. In reviewing the book shortly after its publication, A. H. Copeland wrote, "Posterity may regard this book as one of the major scientific achievements of the first half of the twentieth century."

The May 1958 issue of the *Bulletin of the American Mathematical Society* is devoted to a tribute to John Von Neumann and his work. In their essay on his contributions to the theory of games and mathematical economics, H. W. Kuhn and A. W. Tucker conclude,

> *By his example and through his accomplishments he opened a broad new channel of two-way communication between mathematics and the social sciences. These sciences were fortunate indeed that one of the most creative mathematicians of the twentieth century concerned himself with some of their fundamental problems and constructed strikingly imaginative and stimulating models with which to attack their problems quantitatively. At the same time, mathematics received a vital infusion of fresh ideas and methods that will continue to be highly productive for many years to come. . . . There is a great challenge for other mathematicians to follow his lead in grappling with complex systems in many areas of the sciences where mathematics has not yet penetrated deeply.*

During the 1930s hundreds of prominent German and Austrian scholars fled their native lands to escape the growing oppression of the Nazi movement. Among this group of exiles—which included Von Neumann, Einstein, and Sigmund Freud—was the economist Oskar Morgenstern.

Morgenstern was born in Goerlitz in the German state of Silesia on January 24, 1902. His roots in Germany were well established; one of his ancestors had been a professor of canon law in Leipzig and published a book of sermons in 1508. Morgenstern's father was a poor businessman, and his mother was the illegitimate granddaughter of Emperor Frederick III of Germany. Morgenstern received his secondary and university education in Vienna, earning a doctorate from the University of Vienna in 1925. He returned to the university as a faculty member 4 years later after an extended period of study in London, Paris, and Rome and at Harvard and Columbia Universities in the United States. For nearly a decade, he taught economics in Vienna, edited an academic journal, conducted research, and advised various state agencies. He was the director of the Austrian Institute for Business Cycle Research and served as a consultant to the Austrian National Bank and the Ministry of Commerce. In 1936 he was named a member of the Committee of Statistical Experts of the League of Nations, a position he held until the League was replaced by the United Nations in 1945, and a position from which he helped author a study on *Economic Stability in the Postwar World*.

Morgenstern came to the United States permanently in 1938 when he began a long association with Princeton University. He served on the faculty for 32 years, directed the university's Econometric Research Program and was co-editor of the Princeton Series on Mathematical Economics. He advised the Atomic Energy Commission, the White House, NASA, Congress, and the Rand Corporation. Upon retirement from Princeton in 1970, he accepted a position as professor of economics at New York University. Morgenstern died of cancer at his home in Princeton on July 26, 1977.

Although best known for his collaboration with Von Neumann on game theory, Morgenstern also made many contributions to the theory of business cycles, monetary

policy, international trade, mathematical economics, econometrics, problems of defense strategy, and statistical decision theory. In addition to numerous technical articles and books, he wrote many essays and reviews for such general circulation magazines as *Fortune*, *Scientific American*, *New York Times Magazine*, and *Encounter*.

In an article published in *Fortune*, Morgenstern issued a warning about the uncritical acceptance of imperfect statistics in business, politics, and economics. This essay is highly recommended reading for mathematical modelers who need to construct and test their models from real-world data. In it, Morgenstern notes,

> *Although the natural sciences—sometimes called the "exact" sciences—have been concerned with the accuracy of measurements and observations from their earliest beginnings, they nevertheless suffered a great crisis when it became clear that absolute precision and certainty of important kinds of observations were impossible to achieve in principle. At least all sources of error that occur in the natural sciences also occur in the social sciences: or, in other words, the statistical problems of the social sciences cannot possibly be less serious than those of the natural sciences. But the social sciences pay far less attention to errors than the physical. This is undoubtedly one of the reasons why the social sciences have had a rather uncertain development.*[1]

EXERCISES

I. Introduction

1. Why is it reasonable to conclude that the student sees a greater difference between x and y than between y and z if she prefers y to an even gamble between x and z?

II. Gambles

2. Show that Theorem 1 implies that xIy if and only if $u(x) = u(y)$.

3. Prove condition (1) of Theorem 1.

4. Prove condition (2) of Theorem 1.

5. Prove condition (3) of Theorem 1.

6. Prove condition (5) of Theorem 1.

7. Let p and q be numbers between 0 and 1. Show that the compound gamble $px + (1 - p)[(qy + (1 - q)z)]$ is equivalent to the gamble with three outcomes $px + (1 - p)qy + (1 - p)(1 - q)z$.

III. Axioms of Utility Theory

8. Do Axioms 1–6 seem reasonable to you? If not, which ones would you modify? How would you test the validity of these axioms experimentally?

9. Does Axiom 5 follow from Axiom 4 and the other assumptions about gambles?

10. In the proof of Theorem 2, it is claimed that the gamble $px + (1 - p)y$ is equivalent to the gamble $qx + (1 - q)z$. Why is this true?

11. Suppose a student prefers x to y and y to z and indicates that her difference in happiness between x and y is the same as the difference between y and z. Can you show that she is indifferent between the outcome y and the gamble $.5x + .5z$?

IV. Existence and Uniqueness of Utility

12. If S is a *finite* set of outcomes, show that the proof of Theorem 4 can be considerably simplified.

13. Suppose S is an infinite set where x_0 is the least preferred outcome and x_1 the most preferred one. Discuss how the proof of Theorem 4 can be simplified.

14. Show that the uniqueness of p promised in Theorem 3 is essential to the establishment of a utility function.

15. Show that conditions (1) and (2) of Theorem 4 hold in each of the following cases:

[1] Oskar Morgenstern, "Qui numerare incipit errare incipit," *Fortune* (October 1963), pp. 142–144, 173–174, 178–180.

(a) x and y are in Case (d).

(b) x and y are in Case (e).

(c) x is in Case (d), y is in Case (c).

(d) x is in Case (c), y is in Case (e).

16. Consider an alternative proof of Theorem 4 that proceeds as follows: Define a utility function u whose domain is S in the same manner as the given proof. Then define the utility of any gamble, $px + (1 - p)y$, with outcomes x and y in S, as $u(px + (1 - p)y) = pu(x) + (1 - p)u(y)$. What remains to be proved? Complete the proof.

17. Let T be the set $T = \{0, 1, 2\}$ and let L be the function $L(0) = -17$, $L(1) = \sqrt{23}$, and $L(2) = 10$. Is L a positive linear transformation?

18. Suppose L is a positive linear transformation on a set T. If s and t are distinct elements of T,

(a) Show that $L(s) \neq L(t)$.

(b) If $L(s) = a$ and $L(t) = b$, determine α and β.

19. Verify that Theorem 6 is true in Cases (a), (b), and (e).

20. A student prefers x to y and y to z and finds that $yI[.3x + .7z]$. Construct a utility function consistent with the given information.

21. Suppose a student prefers x to y, y to z, and z to w. Furthermore, $yI[.4x + .6z]$, $yI[.3x + .7w]$, and $zI[.5y + .5w]$. Can you construct a utility function consistent with these observations?

22. If Alexander's utility measures for tuna fish, hamburger, and peanut butter are 60, 48, and 30, respectively, find the gamble with outcomes of peanut butter and tuna fish that he finds indifferent to hamburger.

V. Classification of Scales

23. In their textbook on mathematical psychology, Coombs, Dawes, and Tversky discuss the problem of measurement on a set when, in addition to stating a preference order, you are also able to order differences between alternatives with respect to preferences. They note that "an admissible transformation in this case must preserve not only the order of the scale values but the order of differences between scale values." They then claim that only positive linear transformations preserve the ordering of intervals for any set of objects. Is this true? Can you prove it?

24. In his work on measurement theory, Campbell (see Chapter 7) claimed that only extensive properties can be measured on an interval scale. Since most psychological and sociological properties are intensive, he believed that interval measurement is not possible in these social sciences. Is this argument valid? Is temperature an extensive property?

VI. Interpersonal Comparison of Utility

25. Solve the Registrar's Problem if there are three students (Ann, Joan, and Kathy), three courses (Plate Tectonics, Computer Methods, and Relativity Theory) and each course is restricted to one student. The original utility scales (not normalized) are given in the following table:

	Ann	Joan	Kathy
Plate Tectonics	3	120	–6
Computer Methods	7	90	1.2
Relativity Theory	2	40	5.6

26. Solve the Registrar's Problem of Exercise 25 if each course is open to two students.

27. What would the details of Isbell's argument look like if he were trying to convince you that there is a lower bound to the nonnormalized utility function?

28. Let G be the set of all gambles with outcomes in the finite set S. Prove that any utility function defined on G is necessarily bounded.

SUGGESTED PROJECTS

1. Is there any explanation, in terms of utility, for why people buy insurance policies and lottery tickets?

2. The St. Petersburg Paradox is often cited as an argument that the utility of money is not directly proportional to the amount of money. The paradox concerns a game in which you toss a coin until it lands tails. If this happens for the first time on the nth toss you receive $\$2^n$. How much money are you willing to pay to enter this game? Show that the expected value of your winnings is infinite. Thus, you should be willing to pay any amount to enter.

The "paradox" arises because most people would not pay very much to enter. If you pay $100 to enter the

game and flip "tails" on the first toss, you lose $98! The reluctance of people to enter the game if it has a high entrance fee is due, some believe, to the fact that the value of money is not proportional to the money.

As one way to resolve the paradox, it has been suggested that the utility of money obeys the relation $u(x) = \sqrt{x}$, where x is the number of dollars. Show that if this is true, then the expected utility of the St. Petersburg gamble is finite.

Show, however, that the payoffs in the coin toss can be arranged in such a fashion (say $\$2^{2n}$ instead of $\$2^n$) so that even if $u(x) = \sqrt{x}$, the expected utility is infinite.

Does there exist some function $u(x)$, that increases monotonically with x but for which the St. Petersburg game always has finite expected utility, no matter how the payoffs are arranged? Show that the "paradox" disappears if expected utility is finite and we assume that the rational person acts to maximize expected utility. Is this last assumption valid?

3. Not all social scientists have accepted the axioms of utility theory presented in this chapter. Maurice Allais, Nobel Economics prize winner in 1988, presents a pair of decision situations each involving two gambles. In situation 1, you must choose between

Gamble 1: $500,000 with probability 1

and

Gamble 2: $2,500,000 with probability .1
$500,000 with probability .89
$0 with probability .01

In situation II, the choice is between

Gamble 3: $500,000 with probability .11
$0 with probability .89

and

Gamble 4: $2,500,000 with probability .1
$0 with probability .9

Allais argues that most people prefer gamble 1 to gamble 2 and gamble 4 to gamble 3. Show that these preferences, under the axioms of utility theory, lead to the inconsistent inequalities

(a) $.11\,u(.\$500{,}000) > .1\,u(\$2{,}500{,}000) + .01\,u(0)$

and

(b) $.1\,u(\$2{,}500{,}000) + .01\,u(\$0) > .11\,u(\$500{,}000).$

Do you agree with Allais's preferences? If, after some reflection, you still do, you may wish to read how Leonard Savage reacts to them. What other objections could be raised about the relevance of utility theory as a model of the real world?

4. Almost all theories of social justice and many important societal decisions are based on an implicit comparison of interpersonal utilities. The suggested normalization of Section VI is an explicit way of constructing such comparisons. What principles of equality and fairness are consistent with such a normalization? What principles are violated? Investigate other ways of determining such comparisons.

You can find a listing of references and suggestions for additional reading on the book's website, www.wiley.com/college/olinick

CHAPTER 9 Equilibrium in an Exchange Economy

> The basis of political economy is non-interference. The only safe rule is found in the self-adjusting meter of demand and supply. Do not legislate. Meddle, and you snap the sinews with your sumptuary laws.
>
> —Ralph Waldo Emerson

I. Introduction

We continue our exploration of axiomatic models in this chapter by investigating a model of an exchange economy in which consumers trade goods and services with each other. Each consumer enters the marketplace with an original endowment of commodities and personal desires for more or less of the available items that may be traded.

The principal questions we ask are

- Can we satisfy each consumer's demand with the available supplies?
- What assumptions do we need to make about consumers, goods, and services to guarantee there is some mechanism for matching supply and demand?
- Are simple natural assumptions about consumers inconsistent with each other so that no such mechanism is possible?

We begin with the simplest interesting case: an economy with two consumers and two commodities. We then turn to a more realistic situation with a large but unspecified number of consumers and commodities. We make some reasonable assumptions about the consumers and then investigate whether these assumptions are consistent and whether they necessarily lead to a set of prices for the commodities under which the available supply is adequate for the total demand of each good or service.

II. A Two-Person Economy with Two Commodities

Zoey and Sydney are sisters who both like chocolate and macaroni, but to different degrees. Each initially has a certain amount of both products. They are interested in swapping some of their initial holdings to get a different mixture of chocolate and macaroni.

Suppose that the sisters together have 4 pounds of chocolate and 5 pounds of macaroni. If at any moment Zoey has c pounds of chocolate and m pounds of macaroni, then $0 \leq c \leq 4$

and $0 \leq m \leq 5$. Sydney would have $4 - c$ pounds of chocolate and $5 - m$ pounds of macaroni. Are there redistributions of the chocolate and macaroni on which they both can agree?

A. The Edgeworth Box

We can represent Zoey's bundle of goods by a point P in a rectangle of width 4 and height 5. The point would have coordinates (c, m) in a standard Cartesian coordinate system. Fig. 9.1 shows such a rectangle.

Assuming that the chocolate and macaroni can be divided into arbitrarily small amounts, then any point in this rectangle represents a possible allocation to Zoey. Note that the point $Q(c', m')$ in Fig. 9.1 has $c' > c$ and $m < m'$. Point Q represents an allocation to Zoey where she gets more chocolate, but less macaroni than she gets at point P. From Zoey's perspective horizontal moves to the right and vertical moves up represent increases in each of the two goods.

We don't want to slight Sydney in this treatment. We can use a device called an *Edgeworth box* to represent the allocations to both sisters. The Edgeworth box is a rectangular diagram with Zoey's origin in the lower left corner and Sydney's origin on the diagonally opposite corner. The width of the box is the total amount of chocolate, and the height is the total amount of the macaroni. For Sydney, horizontal shifts to the left represent greater amounts of chocolate and vertical shifts downward mean more macaroni.

The Irish political economist and philosopher Francis Ysidro Edgeworth (1845–1926) introduced this type of diagram in his 1881 book *Mathematical Psychics: An Essay on the Application of Mathematics to the Moral Sciences*. In 1906, Vilfredo Pareto (1848–1923) developed the idea further in *Manual of Political Economy*. Many economics text use the term Edgeworth-Bowley box in honor of Arthur Bowley (1869–1957) who popularized the representation in *The Mathematical Groundwork of Economics* (1924).

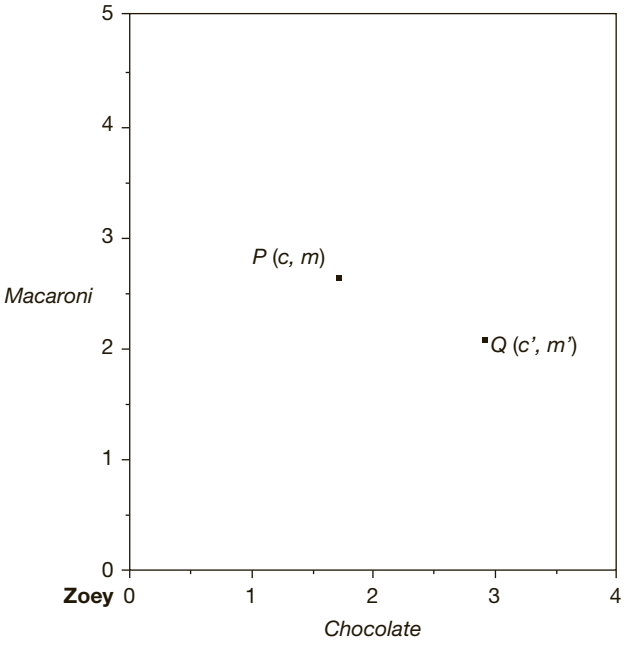


FIGURE 9.1 Possible allocations of chocolate and macaroni to Zoey.

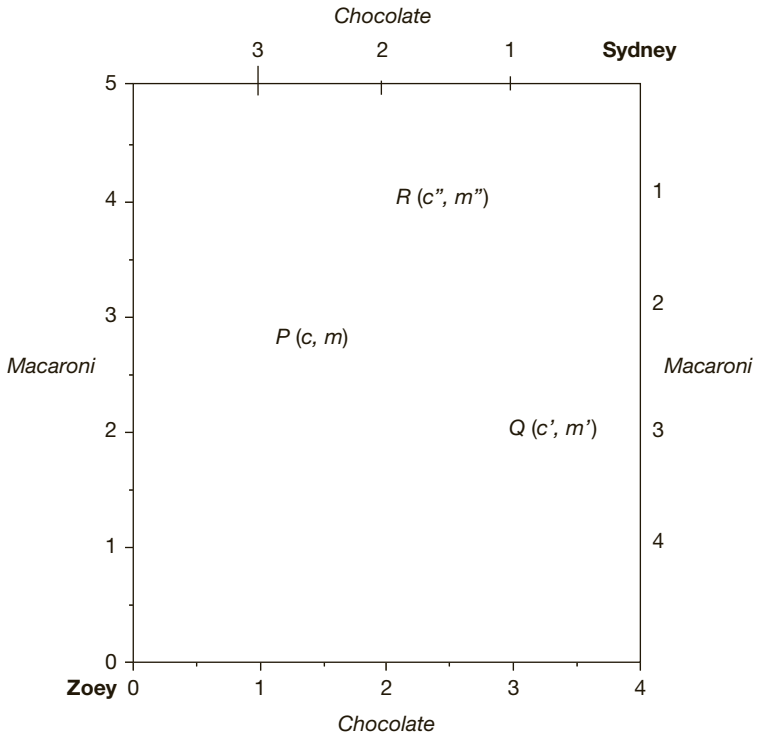


FIGURE 9.2 An Edgeworth box showing possible allocations of chocolate and macaroni to Zoey and Sydney.

Examining Fig. 9.2, it seems reasonable to assume that Zoey would prefer the allocation R to P since she gets more chocolate $(c'' > c)$ and more macaroni $(m'' > m)$. This assumption has a formal name: *consumer insatiability*; informally, it means "more is better." Each consumer prefers any allocation that gives her more of every commodity. [Sydney, we may guess, has the opposite preference between R and P since she winds up with less of both foods.]

B. Indifference Curves

What about Zoey's preference between P and Q? Allocation P gives her more macaroni, but less chocolate than Q. It's not obvious which one she would like better. We would have to ask her. Our second assumption about consumers is that they have preferences. If we ask them to choose between a pair of allocations, they can each tell us which they prefer to the other or if they are indifferent between the two.

We will use notation $P \sim Q$ to indicate indifference between the allocations P and Q. For each allocation P, there will be a set of other feasible allocations all of which Zoey likes equally. The collection of points corresponding to this set is called an *indifference curve*. In Fig. 9.3, we indicate several indifference curves for Zoey. Zoey is indifferent among all the allocations on I1 and indifferent among all the allocations on I3, but she prefers any allocation on I1 to any on I3 (by the principle of consumer insatiability).

In the context of Chapter 8 on utility, it's useful to think that Zoey has a utility function on the set of feasible allocations so that she prefers one allocation to another if it

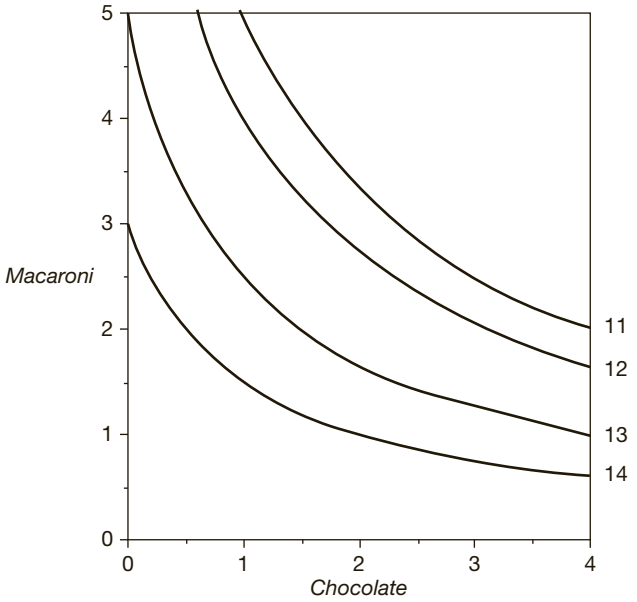


FIGURE 9.3 Some indifference curves for Zoey.

has a higher utility value and is indifferent if the utility of one equals the utility of the other. The indifference curves are then the level curves of her utility function.

If, for example, Zoey's utility for x pounds of chocolate and y pounds of macaroni is given by $u_{\text{Zoey}}(x, y) = (x+1)^2 y^3$, then her indifference curves would have the form $y = \sqrt[3]{\dfrac{a}{(x+1)^2}}$ for different constants a. The utility she would attach to an allocation of 1 pound of chocolate and 4 pounds of macaroni would be $u_{\text{Zoey}}(1, 4) = 256$. The corresponding indifference curve would be $y = \sqrt[3]{\dfrac{256}{(x+1)^2}}$.

Since the total amounts of chocolate and macaroni are fixed, any allocation (c, m) to Zoey determines the allocation $(4 - c, 5 - m)$ to Sydney. For convenience, we will usually write equations for the indifference curves for both sisters in terms of the allocation to Zoey.

Sydney also has preferences among the different possible distributions of chocolate and macaroni. Fig. 9.4 shows some possible indifference curves for Sydney. For each of these curves, any point to the southwest (below and to the left of the curve) is a better distribution for Sydney. Any point to the northeast (above and to the right of the curve) is a worse allocation, exactly the opposite scenario to the one for Zoey.

C. The Bargaining Space

What advice can we give the two sisters about negotiating a redistribution of their original amounts that will please both of them? Suppose, for example, that Zoey comes to the bargaining table with an initial holding (called her *endowment*) of 1 pound of chocolate and 4 pounds of macaroni. The first step is to construct the indifference curve *ZI* for Zoey that runs through this point and the corresponding indifference curve *SI* for Sydney. Fig. 9.5 shows such a possible pair of such indifference curves intersecting at (1, 4).

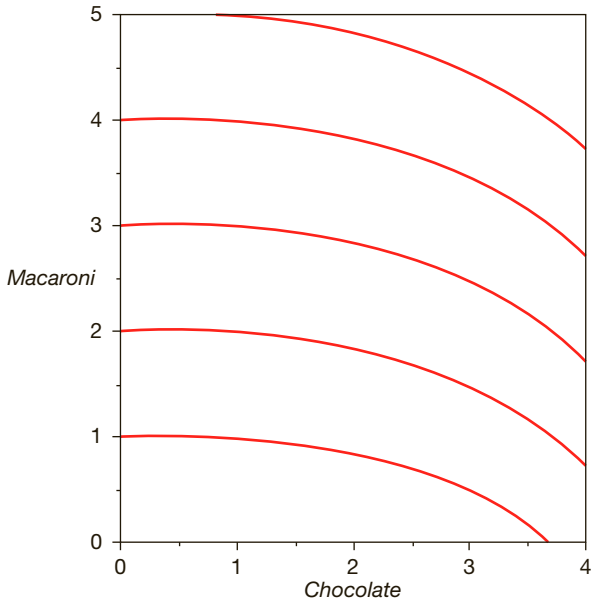


FIGURE 9.4 Some indifference curves for Sydney.

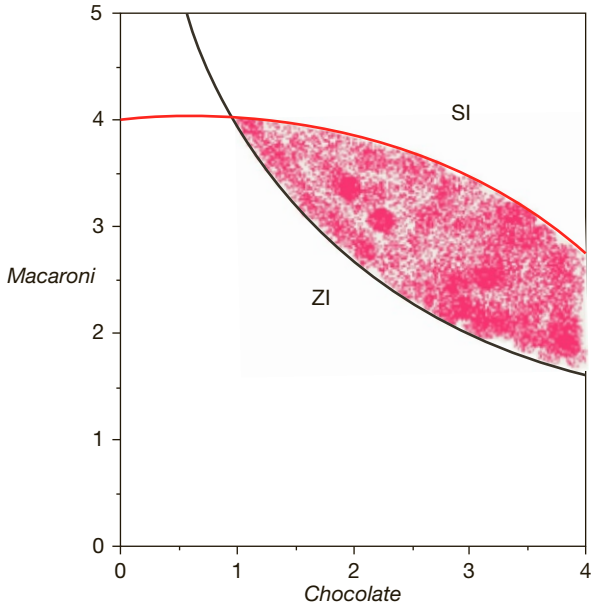


FIGURE 9.5 The indifference curves for Zoey and Sydney passing through the (1, 4) allocation to Zoey. The shaded area represents the bargaining space.

Now Zoey is not going to agree to any redistribution of chocolate and macaroni that gives her less utility than she gets from her initial endowment. She can simply refuse to negotiate and still maintain that utility. She might be willing, however, to consider accepting an allocation at some other point on the indifference curve *ZI*, and she would prefer any suggested redistribution that was above this curve.

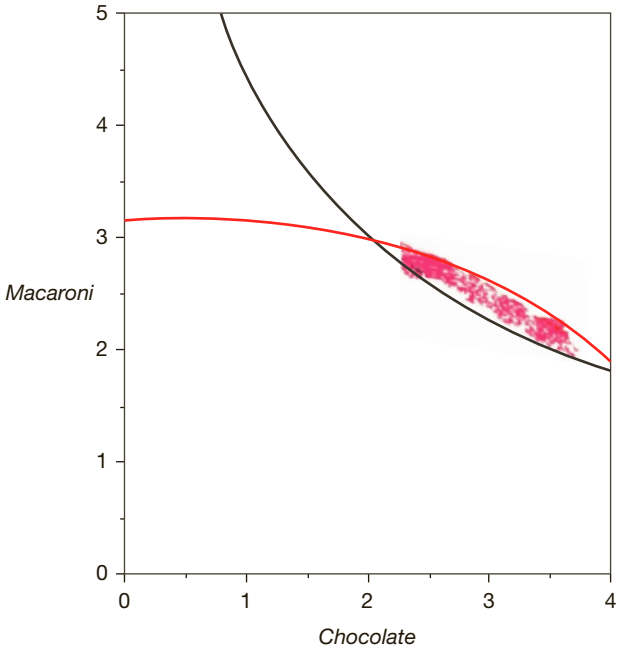


FIGURE 9.6 The indifference curves for Zoey and Sydney passing through the (2, 3) allocation to Zoey. The shaded area represents the bargaining space.

Sydney is in a similar situation; she won't take seriously any offer that represents a point above her indifference curve *SI*. She could agree on another point along *SI* and would be eager to move to a point below that curve. Both would accept a distribution of chocolate and macaroni represented by a point in the area between the curves *ZI* and *SI*—that is, a point that lies above *ZI* and below *SI*. Fig. 9.5 shows a typical situation.

We see in Fig. 9.5 that there is an entire region of points inside our rectangle that lie about Zoey's indifference curve and below Sydney's. We might call this region the *bargaining space*. Both Zoey and Sydney prefer any point in this bargaining space to the status quo point (1, 4).

One such point is (2, 3) representing an allocation of 2 pounds of chocolate to each sister, 3 pounds of macaroni to Zoey and 2 pounds to Sydney. Again, we can graph the indifference curves for both of them that pass through this point. Fig. 9.6 shows that again there is some region between the two curves; the bargaining space has shrunk, but there are still possible distributions that would make each of our consumers even happier than the (2, 3) mixture.

D. Pareto Solutions

Eventually Zoey and Sydney may hit upon redistribution such as (3, 2.16) where the indifference curves intersect at a single point *S* where the curves are tangent to one another as in Fig. 9.7. At such a point, neither sister can do better without making the distribution worse for the other. An allocation of resources in which it is impossible to make any one person better off without making at least one other person worse off is variously called *Pareto-optimal*, *Pareto-efficient*, or a *Pareto solution*. Pareto solutions, then, have a form of equilibrium or stability. Zoey knows she cannot get Sydney to agree to any other allocation that

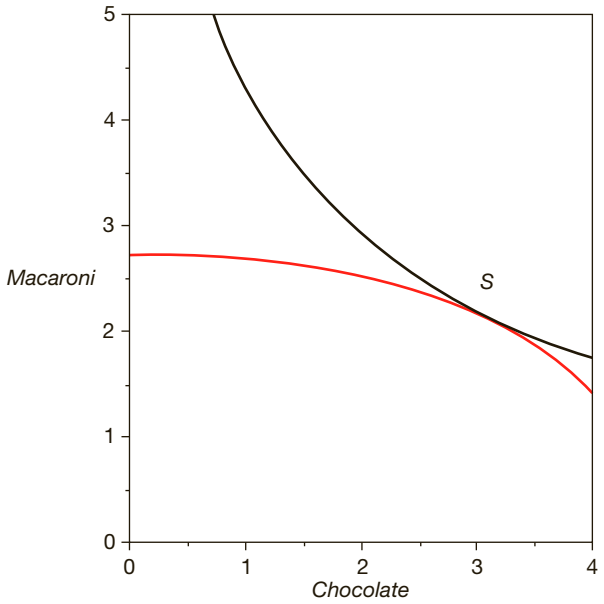


FIGURE 9.7 A Pareto solution for Zoey and Sydney. The indifference curves are tangent to each other at their point of intersection.

makes Zoey happier because that would make Sydney less happy. Sydney realizes the counterpart; she can't hope that Zoey would accept a distribution of greater utility to Sydney.

There may in fact be more than one such point. In our example, let x and y represent the number of pounds of chocolate and macaroni Zoey would receive in a proposed division of the foods. Suppose the indifference curves for Zoey has the form $y = f(x) = \frac{a}{1+x}$, while the indifference curve for Sydney has the equation $y = g(x) = b - \frac{x^3}{50}$ (here a and b are constants). Consider the allocation to Zoey that has $x = 3$ and $y = 2.16$. Then we have $2.16 = \frac{a}{1+3} = \frac{a}{4}$, so $a = 8.64$. We also have $2.16 = b - \frac{3^3}{50} = b - \frac{27}{50} = b - .54$, and hence $b = 2.7$. The derivative of Sydney's indifference curve is $g'(x) = \frac{-3x^2}{50}$; thus, the slope of the tangent line to Sydney's curve at $x = 3$ is $-\frac{27}{50} = -.54$. Since Zoey's curve has $f(x) = \frac{8.64}{1+x}$, we have $f'(x) = -\frac{8.64}{(1+x)^2}$. Thus, $f'(3) = -\frac{8.64}{(1+3)^2} = -\frac{8.64}{16} = -.54$. We see that Zoey's and Sydney's indifference curves have equal tangents at the intersection point (3, 2.16). Thus, (3, 2.16) is a Pareto solution for this example.

In this case, any point in the bargaining space along the curve $y = h(x) = \frac{3x^2(1+x)}{50}$ is a possible *Pareto solution*. To see why this claim is true, suppose (k, c) is a point in the bargaining space where $c = h(k) = \frac{3k^2(1+k)}{50}$. Since Sydney's indifference curves have the form $y = g(x) = b - \frac{x^3}{50}$, we need to choose $b = c + \frac{k^3}{50} = \frac{3k^2(1+k)}{50} + \frac{k^3}{50} = \frac{3k^2+4k^3}{50}$ to

single out the indifference curve passing through the point (k, c). The slope of the tangent line to Sydney's indifference curve at this point is $g'(k) = \dfrac{-3k^2}{50}$. The indifference curve for Zoey that passes through (k, c) must have $c = \dfrac{3k^2(1+k)}{50} = f(k) = \dfrac{a}{1+k}$. Thus, $a = \dfrac{3k^2(1+k)^2}{50}$ and so $f'(x) = -\dfrac{a}{(1+x)^2} = -\dfrac{3k^2(1+k)^2}{50(1+x)^2}$. Hence, the slope of the tangent line to Zoey's indifference curve at the point (k, c) is $f'(k) = -\dfrac{a}{(1+k)^2} = -\dfrac{3k^2(1+k)^2}{50(1+k)^2} = -\dfrac{3k^2}{50}$. Thus, the indifference curves for Zoey and Sydney are tangent to each other at the point of intersection. The points in the bargaining space along the curve $y = h(x)$ is called the *contract curve*.

How did we find a formula for the contract curve in this case? Let $h(x)$ be our unknown contract curve function with $(k, c) = (k, h(k))$ a point in the bargaining space. We need both indifference curves to pass through this point with identical derivatives. For Zoey's curve to pass through $(k, h(k))$, we need $h(k) = \dfrac{a}{1+k}$ so $a = (1+k)h(k)$, which gives $f(x) = \dfrac{(1+k)h(k)}{1+x}$. This relationship implies $f'(x) = -\dfrac{(1+k)h(k)}{(1+x)^2}$ so that $f'(k) = -\dfrac{(1+k)h(k)}{(1+k)^2} = \dfrac{-h(k)}{1+k}$. The derivative for Sydney's curve will be $g'(k) = \dfrac{-3k^2}{50}$. For the derivatives to be equal, we need $\dfrac{-3k^2}{50} = \dfrac{-h(k)}{1+k}$, and hence, we see that $h(k) = \dfrac{3k^2(1+k)}{50}$.

Fig. 9.8 shows a pair of indifference curves for Zoey and Sydney, the solution space and the contract curve. We can expect that the two sisters will eventually agree to accept some point along the contract curve, but we cannot predict which one.

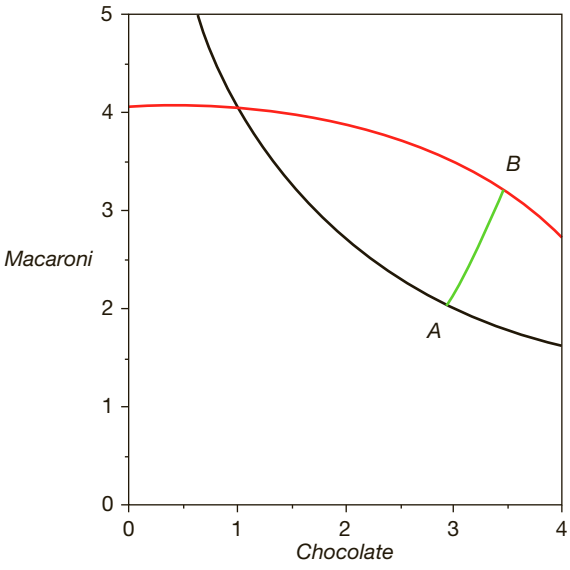


FIGURE 9.8 The curve running from A to B is the contract curve for Zoey and Sydney.

In this case, with two consumers and two commodities, we were able to find some allocations to our pair of economic agents, which have stable equilibrium properties. Our analysis was based on knowing specific equations for their indifference curves. It is certainly not clear that given two arbitrary utility functions whether there will always be at least one point where the curves are tangent to each other.

Rather than pursue here models and theorems about two-person, two-commodity economies, we will look at the question of equilibrium in an exchange economy where there are potentially millions of consumers and thousands of commodities.

III. An *m*-Person Economy

Let's move on to a more realistic economy with many agents and a larger number of commodities. We will restrict ourselves, however, to an *exchange economy*—that is, one in which there is no production. The basic question we address is whether there is a set of prices under which total demand equals total supply for all commodities.

Here is a summary view of the model: the total *supply* is owned by individual *consumers* who are willing to exchange some of their initial holdings of commodities (*endowments*) for *commodities* owned by others. Under a particular set of prices, each consumer has a certain *wealth*, the value of the consumer's endowment. That wealth will determine what the consumer can afford—so each consumer has a *budget constraint*. Working with that constraint, we assume that the consumer has, most desired collection of commodities, called a commodity bundle. Summing all these commodity bundles, we can determine how the *total demand* compares with the total supply.

We need to define many of the terms in this summary and discuss how we can represent them mathematically. These include:

Commodities	Total Supply	Prices
Consumers	Endowment	Wealth
Budget Constraint	Demand Function	Total Demand

A. Commodities

A *commodity* is a *good* or *service* whose characteristics are precisely and completely described, including the location at which it is available and the time at which it is available. Goods include such items as food staples, household furnishings, automobiles, and other tangible items. Services refer to less tangible items that still satisfy individuals' wants and needs, such as medical operations, apartment rentals, haircuts, and the like.

We assume that each commodity has a unit of measurement but is arbitrarily divisible into smaller quantities. A complete description of a commodity would include its location and time of availability. Thus, a DVD of a particular film available in Chicago at noon next Monday would be considered as a different commodity than a DVD of the same film available at the same spot in Chicago two months from now. There will be a large, but finite number of commodities that we label 1, 2, 3, . . . , h, . . . , l. The price for one unit of commodity h is a nonnegative number we will denote as p_h.

By a *consumption vector* or *commodity bundle*, we mean a vector $\mathbf{x} = (x_1, x_2, \ldots, x_h, \ldots, x_l)$ where x_h is the number of units of commodity h. Note that a commodity bundle is a point in l-dimensional space all of whose components are nonnegative.

We denote the set of all commodity bundles as $\mathbb{R}^{\ell}_{+}$. If $\mathbf{p} = (p_1, p_2, \ldots, p_h, \ldots, p_l)$ is a vector of unit prices, then the *cost* of a consumption vector $\mathbf{x}$ is the dot product of $\mathbf{p}$ and $\mathbf{x}$:

$$p_1x_1 + p_2x_2 + \cdots + p_\ell x_\ell = \sum_{h=1}^{\ell} p_h x_h = \mathbf{p} \bullet \mathbf{x}$$

Example

Suppose $l = 3$ and $\mathbf{p} = (1, 2, 4)$. If three commodity bundles x, y, and z are given as $x = (2, 1, 3)$, $y = (8, 0, 2)$, and $z = (5, 1, 1)$, then the costs of these vectors are

$$\begin{aligned} \text{Cost of } x &= (1, 2, 4) \bullet (2, 1, 3) = 2 + 2 + 12 = 16 \\ \text{Cost of } y &= (1, 2, 4) \bullet (8, 0, 2) = 8 + 0 + 8 = 16 \\ \text{Cost of } z &= (1, 2, 4) \bullet (5, 1, 1) = 5 + 2 + 4 = 11 \end{aligned}$$

Observe that while x and y are different bundles, they cost the same amount.

B. Consumers, Endowments, and Demand Functions

We turn now to a mathematical representation of our economic agents, who we will call *consumers*. For our model, a consumer is characterized by an *endowment* and a *demand function*. We assume that each consumer has an initial vector $\mathbf{e} = (e_1, e_2, \ldots, e_\ell)$ of resources (called his *endowment*), which is what he owns of each commodity before any exchange. Each component e_h of an endowment vector is a nonnegative number. Hence, each endowment vector $\mathbf{e}$ is also a vector in $\mathbb{R}^{\ell}_{+}$.

For a fixed set of prices $\mathbf{p}$, the consumer's *wealth* w is the value of his or her endowment: $w = \mathbf{p} \bullet \mathbf{e} = (e_1p_1 + \cdots + e_hp_h + \cdots + e_lp_l)$.

We imagine the consumers trading their initial bundle of goods and services (their endowments) for other commodity bundles either by a direct swap or, more efficiently, selling their endowment at the current prices and using the funds to purchase a more desired consumption vector.

We assume each consumer has a demand function f whose output is that individual's most desired consumption bundle. Reality dictates that the consumer's most wanted consumption vector will depend on his endowment and the prices. The consumer can't buy a collection of good and services that costs more than he can afford. His wealth depends in turn on both prices and endowment.

Mathematically speaking, a consumer's demand function f is a function whose inputs are a pair of l-dimensional vectors, $\mathbf{p}$ and $\mathbf{e}$, and whose output $f(\mathbf{p}, \mathbf{e})$ is a vector in $\mathbb{R}^{\ell}_{+}$. These demand functions may vary from consumer to consumer. Two consumers with identical wealth, or even identical endowments, may desire very different collections of goods and services they can afford.

As we have mentioned, each consumer is limited, however, by a *budget constraint*. The desired bundle $f(\mathbf{p}, \mathbf{e})$ cannot cost more than the consumer's wealth under the prevailing prices. In mathematical terms, the consumer can choose any $\mathbf{x}$ in $\mathbb{R}^{\ell}_{+}$ such that $\mathbf{p} \bullet \mathbf{x} \leq w = \mathbf{p} \bullet \mathbf{e}$ —that is,

$$\mathbf{p} \bullet f(\mathbf{p}, \mathbf{e}) \leq \mathbf{p} \bullet \mathbf{e}$$

Example

Suppose $\mathbf{e} = (10, 20)$ and $\mathbf{p} = (.7, .3)$. Then the consumer's wealth is $(.7)(10) + (.3)(20) = 7 + 6 = 13$. Then the consumer's budget constraint set consists of all vectors $\mathbf{x} = (x_1, x_2)$ such that $.7x_1 + .3x_2 \leq 13$.

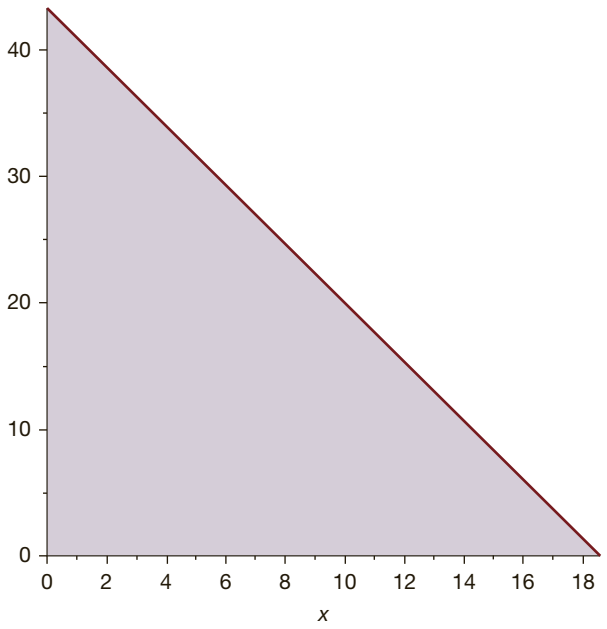


FIGURE 9.9 A consumer's budget constraint in a two-commodity economy.

C. Normalized Prices

So far we have not said anything about what units of currency are used for prices. Suppose that in the last example, the prices were originally quoted in dollars, but there is an agreement to switch to nickels. Now every price is inflated by a factor of 20. Something that sold for \$2 per unit suddenly costs 40 nickels. Each consumer's wealth, however, goes up by the same factor. Thus, our consumer with endowment $\mathbf{e} = (10, 20)$ now has a wealth of $(14)(10) + (6)(20) = 140 + 120 = 260$ nickels. The budget constraint is all vectors $\mathbf{x} = (x_1, x_2)$ such that $14x_1 + 6x_2 \leq 260$. This inequality, however, is equivalent to $20(.7x_1 + .3x_2) \leq (20)(13)$ or $.7x_1 + .3x_2 \leq 13$. Thus, the budget constraint does not change at all!

The same result holds if we switch from dollars to euros to yens or to any other currency. Prices and hence wealth are inflated or deflated by the same factor so there is no change at all in the budget constraint set. We would expect the consumer faced with the constraint set would make the same choice of a commodity bundle no matter what currency is used in the marketplace.

We formalize this assumption about consumers as an axiom of *homogeneity*:

$$f(\lambda \mathbf{p}, \mathbf{e}) = f(\mathbf{p}, \mathbf{e}) \text{ for all } \lambda > 0$$

Since we are free to choose the currency unit, we will usually work with *normalized prices*—that is, prices that sum to 1. Assuming that at least one unit price is positive, we can normalize any price vector by dividing each price by the sum of all the prices. As an example, the price vector $\mathbf{q} = (1, 4, 3, 2)$ is normalized by dividing each component by the sum of the components $(1 + 4 + 3 + 2 = 10)$ to obtain $\mathbf{p} = (.1, .4, .3, .2)$.

Note that the individual unit prices in the normalized form have the same relative weights as the original vector; in our example, for instance, the second price in $\mathbf{q}$ is 4 times the first price and twice the fourth price; the same holds true for the normalized vector $\mathbf{p}$. Thus, we don't lose any vital information by restricting ourselves to normalized prices.

The set $\prod$ of all possible normalized price vectors is the set of l-dimensional vectors whose components are nonnegative numbers summing to 1. More formally,

$$\prod = \{\mathbf{p} : \mathbf{p} \in \mathbb{R}^{\ell}_{+} \text{ and } \sum p_h = 1\}$$

In the simplest case of a two-commodity economy $(l = 2)$, a normalized price vector $\mathbf{p}$ has the form $\mathbf{p} = (p_1, p_2)$ where $p_1 \geq 0$, $p_2 \geq 0$ and $p_1 + p_2 = 1$. Geometrically, the set $\prod$ is the line segment in the plane joining the points (0, 1) and (1, 0). If $l = 3$, then $\prod$ is all the points on or inside equilateral triangle in 3-space with vertices (1, 0, 0), (0, 1, 0), and (0, 0, 1). See Fig. 9.10.

Our second major assumption (beyond homogeneity) about demand functions is called *consumer insatiability*. It simply states that if a consumer can afford to acquire more commodities, then he will choose to do so. Faced with choosing between two commodity bundles x and y both lying in the budget constraint set with $x_h \geq y_h$ for all commodities h and $x_k > y_k$ for at least one commodity k, the consumer will prefer x to y. The bundle x does more to meet the consumer's wants or needs and hence provides more happiness or utility than bundle y.

Consumer insatiability may at first sight seem unreasonable. If I already have enough milk in a particular bundle x that I can consume before it spoils, why would I want a bundle y with even more milk even though I can afford y? Recall, however, that two gallons of milk from my favorite local grocery are considered to be different commodities if they are available at different times? Thus, I might spend some of my wealth left over after purchasing x to add to the bundle milk from that grocer that I will pick up next week. As an alternative, we could consider money in a savings account to be a commodity and "spend" the difference

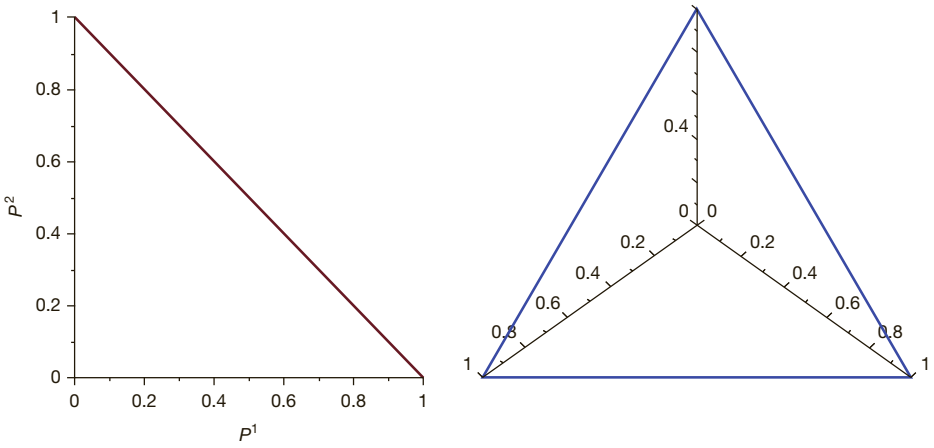


FIGURE 9.10 The set of normalized prices for economies with two commodities (on the left) and three commodities (on the right). For $l = 2$, $\prod$ is the straight-line segment between (0, 1) and (0, 1). For $l = 3$, $\prod$ is the face of the equilateral triangle with vertices at (1, 0, 0), (0, 1, 0), and (0, 0, 1).

between my wealth and the cost of x as a deposit in the bank. Thus, our more sophisticated view of a commodity makes consumer insatiability a more plausible assumption.

Geometrically speaking, if bundle x "lies northeast" of bundle y and both are in the budget constraint set, then our consumer prefers x to y. Recall Fig. 9.2 where Zoey preferred $x = (c'', m'')$ to $y = (c, m)$ because $c'' > c$ and $m'' > m$.

Consumer insatiability leads to the conclusion that the output of a consumer's demand function will be a commodity bundle x whose cost under the current prices $\mathbf{p}$ will equal the consumer's wealth—that is,

$$\mathbf{p} \bullet f(\mathbf{p}, \mathbf{e}) = \mathbf{p} \bullet \mathbf{e}$$

We will make one more reasonable assumption about consumer demand functions: the output of the demand function f will not change very much if there are small changes in prices or endowment. Mathematically, this assumption is the assertion that consumer demand functions are *continuous*.

D. Consumer Interaction

We consider now the interaction of m consumers: $1, 2, \ldots, i, \ldots, m$. We have m demand functions $f_1, f_2, \ldots, f_m$ and m endowment vectors $\mathbf{e}^1, \mathbf{e}^2, \ldots, \mathbf{e}^m$ where $\mathbf{e}^i$ is the endowment of consumer i and f^i is her demand function.

To find the total available amount of some commodity, we would simply add together the corresponding component of each of the endowment vectors. The total supply of commodity 3, for example, would be the sum of the third components of $\mathbf{e}^1, \mathbf{e}^2, \ldots,$ and $\mathbf{e}^m$. Since addition of vectors is defined by adding together corresponding components, the sum of all the endowment vectors gives us a *total supply vector*:

$$\text{Total Supply } \mathbf{S} = \sum_{i=1}^{m} \mathbf{e}^{\mathbf{i}}.$$

The total supply vector $\mathbf{S}$ is an element of $\mathbb{R}^{\ell}_{+}$ whose hth component is the total available amount of commodity h.

In a similar fashion, if we add the desired commodity bundles together we obtain a *total demand vector* $\mathbf{D}$, which is also in $\mathbb{R}^{\ell}_{+}$:

$$\text{Total Demand } \mathbf{D} = \sum_{i=1}^{m} f^i(\mathbf{p}, \mathbf{e}^{\mathbf{i}})$$

Once we have the total supply and total demand vectors, we can create their difference, which defines an *excess demand function* F whose output is the *excess demand vector*:

$$\text{Excess Demand } F(\mathbf{p}) = \sum_{i=1}^{m} f^i(\mathbf{p}, \mathbf{e}^{\mathbf{i}}) - \sum_{i=1}^{m} \mathbf{e}^{\mathbf{i}} = \sum_{i=1}^{m} \left[f^i(\mathbf{p}, \mathbf{e}^{\mathbf{i}}) - \mathbf{e}^{\mathbf{i}} \right]$$

If the hth component of the excess demand vector is positive, then the overall demand for commodity h exceeds the total supply. If that component is zero, then we have exactly as much of this commodity to satisfy everyone's desires. If the component is negative, then

there is more supply than demand for that particular good or service. The vector $F(\mathbf{p})$ is an l-dimensional vector, which may have a mixture of positive, negative, and zero components. We will use the notation $F(\mathbf{p}) \leq \mathbf{0}$ to signify that each component of the excess demand vector is nonpositive (less than or equal to 0).

A price vector $\mathbf{p}$ is called an equilibrium price vector if $F(\mathbf{p}) \leq \mathbf{0}$. Under equilibrium prices, there is enough of every commodity to meet or exceed everyone's demands. A fundamental question for us is: Under what conditions does there exist an equilibrium price vector?

Example

Suppose our economy has $l = 3$ commodities and $m = 4$ consumers with the following endowments:

$\mathbf{e}^1 = (2, 1, 8)$ $\mathbf{e}^2 = (6, 0, 3)$

$\mathbf{e}^3 = (4, 4, 4)$ $\mathbf{e}^4 = (2, 5, 4)$

Then the total supply is (14, 10, 19).

(A) Suppose our prices (before normalization) are given by $\mathbf{p} = (1, 2, 3)$. Then the wealth of our consumers is given by $\mathbf{w} = (w_1, w_2, w_3, w_4) = (28, 15, 24, 24)$. Suppose also that the individual demands under $\mathbf{p}$ are

$f^1(\mathbf{p}) = (12, 5, 2)$	$f^2(\mathbf{p}) = (2, 5, 1)$
$f^3(\mathbf{p}) = (1, 1, 7)$	$f^4(\mathbf{p}) = (9, 6, 1)$

so that the total demand vector is (24, 17, 11). Then the excess demand vector is $(24, 17, 11) - (14, 10, 19) = (10, 7, -8)$. There is an excess demand for commodity 1 and for commodity 2, but an abundant supply of commodity 3. Thus, neither $\mathbf{p} = (1, 2, 3)$, nor its normalized form $(\frac{1}{6}, \frac{1}{3}, \frac{1}{2})$ are equilibrium prices.

(B) Suppose our prices are $\mathbf{q} = (5, 0, 4)$. The wealth vector is (42, 42, 36, 26). If the demand vectors under $\mathbf{q}$ are:

$f^1(\mathbf{q}) = (6, 3, 3)$	$f^2(\mathbf{q}) = (2, 2, 8)$
$f^3(\mathbf{q}) = (2, 0, 13/2)$	$f^4(\mathbf{q}) = (5, 2, 1)$

then the total demand vector is (14, 8, 19). Then the excess demand vector is $F(\mathbf{q}) = (14, 8, 19) - (14, 10, 19) = (0, -2, 0)$ so $\mathbf{q}$ is an equilibrium price vector. Note that the commodity for which we have a surplus (commodity 2) has a price of 0.

(C) Consider a third set of prices $\mathbf{r} = (3, 8, 5)$ where the wealth vector becomes (54, 33, 64, 66). If the demand vectors under $\mathbf{r}$ are:

$f^1(r) = (3, 0, 9)$	$f^2(r) = (3, 3, 0)$
$f^3(\mathbf{p}) = (0, 3, 8)$	$f^4(\mathbf{p}) = (7, 6, 3)$

then the total demand vector is (14, 10, 19). The excess demand vector is $F(\mathbf{r}) = (14, 10, 19) - (14, 10, 19) = (0, 0, 0)$ so $\mathbf{r}$ is an equilibrium price vector. Here total supply and total demand are the same for every commodity.

We can now state in a simple manner the central questions we hope our model will address:

1. Do equilibrium prices $\mathbf{p}$ exist under our assumptions?
2. What extra assumptions must we make or which assumptions must we weaken to guarantee the existence of such a price system $\mathbf{p}$?
3. Is such a $\mathbf{p}$ unique?
4. How do we find or compute such a $\mathbf{p}$?

E. Walras's Law

In this section, we will derive an important consequence of our assumption of consumer insatiability and its implications for equilibrium prices.

Suppose $\mathbf{p}$ is any set of prices, equilibrium or not. Consumer insatiability tells us that

$$\mathbf{p} \bullet f^i(\mathbf{p}, \mathbf{e}^i) = \mathbf{p} \bullet \mathbf{e}^i$$

for each consumer i. Let's rewrite this equality as

$$0 = \mathbf{p} \bullet f^i(\mathbf{p}, \mathbf{e}^i) - \mathbf{p} \bullet \mathbf{e}^i = \mathbf{p} \bullet \left[f^i(\mathbf{p}, \mathbf{e}^i) - \mathbf{e}^i\right]$$

and then sum these numbers over all our consumers to obtain

$$0 = \sum_{i=1}^{m} 0 = \sum_{i=1}^{m} \mathbf{p} \bullet \left[f^i(\mathbf{p}, \mathbf{e}^{\mathbf{i}}) - \mathbf{e}^{\mathbf{i}}\right].$$

Elementary properties of vector arithmetic and the dot product allow us to rewrite the right-hand side of this last equation as

$$0 = \sum_{i=1}^{m} \mathbf{p} \bullet \left[f^i(\mathbf{p}, \mathbf{e}^{\mathbf{i}}) - \mathbf{e}^{\mathbf{i}}\right] = \mathbf{p} \bullet \sum_{i=1}^{m} \left[f^i(\mathbf{p}, \mathbf{e}^{\mathbf{i}}) - \mathbf{e}^{\mathbf{i}}\right] = \mathbf{p} \bullet F(\mathbf{p})$$

where the last equality follows from the definition of the excess demand function. We have derived an important result called Walras's Law:

THEOREM 9.1 Walras's Law For any price vector $\mathbf{p}$, we have $\mathbf{p} \bullet F(\mathbf{p}) = 0$.

This result is named for the French economist Marie Esprit Léon Walras (1834–1910) who expounded it in several of important books *Eléments d'economie politique pure* (1874–1877) and *Théorie mathematique de la richesse sociale* (1883). The English

philosopher John Stuart Mill formulated a similar idea without a mathematical representation in the 1840s. Note that we can also write Walras's Law in a more compact form $\sum_{h=1}^{l} p_h F_h(\mathbf{p}) = 0$.

Walras's Law has an interesting implication if $\mathbf{p}$ is an equilibrium price vector. In this case $F(\mathbf{p}) \leq 0$ so each number $F_h(p)$ is less than or equal to zero. Since prices are nonnegative, each product $p_h F_h(\mathbf{p})$ is also less than or equal to 0. But according to Walras's Law, the sum of these nonpositive numbers $p_h F_h(\mathbf{p})$ is 0. Such a sum can only equal 0 if every term is 0; if any term were negative, the entire sum would be negative.

In particular, if there is a commodity h for which there is an excess supply under an equilibrium price vector $\mathbf{p}$ so that $F_h(\mathbf{p})$ is negative, we must have $p_h = 0$:

THEOREM 9.2 Under equilibrium prices, the price of a commodity that has an excess supply must be zero.

IV. Existence of Economic Equilibrium

In this section, we will establish the main result of this chapter: under our assumptions about consumers, at least one price equilibrium will always exist. We will first present a proof for the $l = 2$ case that only uses results from elementary calculus. Then we provide an argument for the general case that rests on a 20th-century theorem in topology.

We list here the three important assumptions we are making about consumers and their demand functions f:

1. f is continuous; small changes in $\mathbf{p}$ or $\mathbf{e}$ result in small changes in $\mathbf{x}$.
2. $f(\lambda\mathbf{p},\mathbf{e}) = f(\mathbf{p},\mathbf{e})$ for all $\lambda > 0$.
3. the consumer is insatiable—that is, for all $\mathbf{p}$ and $\mathbf{e}$, we have

$$\mathbf{p} \bullet f(\mathbf{p}, \mathbf{e}) = \mathbf{p} \bullet \mathbf{e}$$

A. Price Equilibrium in a Two-Commodity Economy

THEOREM 9.3 Equilibrium prices always exist in a two-commodity economy.

Proof We can describe the set Π of normalized prices in a two-commodity economy as

$$\Pi = \{\mathbf{p} : \mathbf{p} = (p_1, p_2) \text{ where } p_1 \geq 0, p_2 \geq 0, \text{ and } p_1 + p_2 = 1\}$$

Consider two special price structures in $\prod$, $\mathbf{q} = (1,0)$ and $\mathbf{r} = (0,1)$. If either $\mathbf{q}$ or $\mathbf{r}$ is an equilibrium price vector, we are done. So suppose neither one is. Consider $\mathbf{q}$ first. By Walras's Law,

$$0 = \mathbf{q} \bullet F(\mathbf{q}) = q_1 F_1(\mathbf{q}) + q_2 F_2(\mathbf{q}) = 1F_1(\mathbf{q}) + 0F_2(\mathbf{q}) = F_1(\mathbf{q})$$

Thus, $F_1(\mathbf{q}) = 0$. Since $\mathbf{q}$ is not an equilibrium vector, we must have $F_2(\mathbf{q}) > 0$. A similar argument shows that $F_1(\mathbf{r}) > 0$.

Since there are only two prices and they sum to 1, the price vector is determined once we know the price of commodity 1. Thus, we can think of the price vector as a function of its first component.

Take any price vector $\mathbf{p}^*$ "between" $\mathbf{r}$ and $\mathbf{q}$—that is, $\mathbf{p}^* = (p_1^*, p_2^*)$ where $0 < p_1^* < 1$. Such a $\mathbf{p}^*$ will also have p_2^* strictly between 0 and 1. Fig. 9.11 shows the relative positions of $\mathbf{r}$, $\mathbf{q}$, and $\mathbf{p}^*$. If $\mathbf{p}^*$ turns out to be an equilibrium vector, then we are done. What can we say if $\mathbf{p}^*$ is not an equilibrium set of prices?

If we apply Walras's Law to $\mathbf{p}^*$, we have

$$0 = \mathbf{p}^* \bullet F(\mathbf{p}^*) = p_1^* F_1(\mathbf{p}^*) + p_2^* F_2(\mathbf{p}^*)$$

Since p_1^* and p_2^* are both positive, we conclude that $F_1(\mathbf{p}^*)$ and $F_2(\mathbf{p}^*)$ must be of opposite sign. We may assume, without loss of generality, that $F_2(\mathbf{p}^*)$ is negative. [In Exercise 20, you will examine the case that $F_1(\mathbf{p}^*) < 0$.] Now we have $F_2(\mathbf{q}) > 0$ and $F_2(\mathbf{p}^*) < 0$.

As we noted above, in a two-commodity economy, F_2 is really a function of the first component of a price vector. Thus, F_2 is a continuous real-valued function on the closed interval [0, 1]. Since $\mathbf{q} = (1,0)$ and $\mathbf{p}^* = (p_1^*, p_2^*)$ with $0 < p_1^* < 1$, we have $F_2(p_1^*) < 0 < F_2(1)$; we can also think of this inequality as $F_2(\mathbf{p}^*) < 0 < F_2(\mathbf{q})$. By the Intermediate Value Theorem of elementary calculus, there is at least number s_1 between p_1^* and 1 with $F_2(s_1) = 0$. Since s_1 determines a price vector $\mathbf{s} = (s_1, 1 - s_1)$, we have $F_2(\mathbf{s}) = 0$.

Now apply Walras's Law to the price vector $\mathbf{s}$:

$$0 = \mathbf{s} \bullet F(\mathbf{s}) = s_1 F_1(\mathbf{s}) + s_2 F_2(\mathbf{s}) = s_1 F_1(s) + s_2(0) = s_1 F_1(\mathbf{s})$$

We have $s_1 F_1(\mathbf{s}) = 0$ with s_1 strictly positive; hence, $F_1(\mathbf{s})$ must also be 0. Since $F_1(\mathbf{s}) = F_2(\mathbf{s}) = 0$, we have $F(\mathbf{s}) = 0$. Be definition, $\mathbf{s}$ is an equilibrium price vector. We have shown the existence of equilibrium prices in a two-commodity economy. ◇

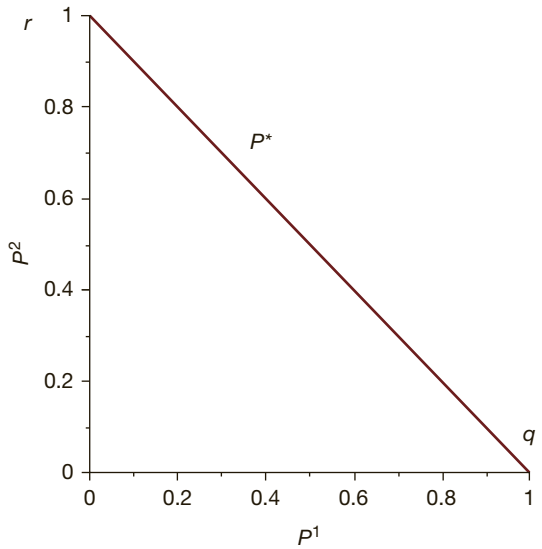


FIGURE 9.11 The relative position of $\mathbf{r}$, $\mathbf{p}^*$, and $\mathbf{q}^*$.

B. Price-Setting Agency

Let's turn to the more general case of an economy with large numbers of commodities and consumers. We will demonstrate that our assumptions about our consumers are strong enough to guarantee the existence of an equilibrium price vector.

Imagine that there is an official Price-Setting agency that is trying to establish an equilibrium set of prices. The agency might announce a tentative set $\mathbf{p}$ of normalized prices and then poll all the consumers to obtain their desired commodity bundles. The agency would then compare total demand and total supply for all the commodities in the market. The agency would expect that an arbitrarily chosen $\mathbf{p}$ would result in excess demands for some goods and services and an oversupply of others.

The agency might then consider revising the prices, hoping to move toward equilibrium by raising the price on commodities with too great a demand. Suppose that under $\mathbf{p}$, there is not enough supply of commodity h to meet the total desired by the aggregate of consumers. Since the endowments are fixed, it makes sense to raise the price of commodity h by some positive amount δ_h to lessen the demand. How big should δ_h be? We need a number that we know is nonnegative and related to commodity h. One simple answer is to take δ_h to be the larger of 0 and $F_h(\mathbf{p})$:

$$\delta_h(\mathbf{p}) = \max(0, F_h(\mathbf{p})) = \begin{cases} 0 & \text{if } F_h(\mathbf{p}) \leq 0 \\ F_h(\mathbf{p}) & \text{if } F_h(\mathbf{p}) > 0 \end{cases}$$

If we compute such a quantity for each commodity, we can form a vector of price modifications $\delta(\mathbf{p}) = (\delta_1(\mathbf{p}), \delta_2(\mathbf{p}), \ldots, \delta_h(\mathbf{p}), \ldots, \delta_l(\mathbf{p}))$ and a new price vector $\mathbf{p} + \delta(\mathbf{p})$. We need to normalize this vector by dividing each component by the sum of all the components. Let $a = \mathbf{u} \bullet (\mathbf{p} + \delta(\mathbf{p}))$ where $\mathbf{u} = (1,1,\ldots,1,\ldots,1)$. Finally we define the transformed prices as

$$T(\mathbf{p}) = (1/a)(\mathbf{p} + \delta(\mathbf{p}))$$

Example

Suppose $\mathbf{p} = \left(\frac{1}{6}, \frac{2}{6}, \frac{3}{6}\right)$ and $F(\mathbf{p})$ turns out to be

$$F(\mathbf{p}) = \text{Demand} - \text{Supply} = (24,17,11) - (14,10,19) = (10,7,-8)$$

Then $\delta_1 = 10$, $\delta_2 = 7$, $\delta_3 = 0$ so that the new prices are initially

$$\left(\frac{1}{6} + 10, \frac{2}{6} + 7, \frac{3}{6} + 0\right) = \left(\frac{61}{6}, \frac{44}{6}, \frac{3}{6}\right)$$

which sum to $\frac{61 + 44 + 3}{6} = \frac{108}{6}$. We normalize by multiplying by $\frac{6}{108}$ to obtain

$$T(\mathbf{p}) = \left(\frac{61}{108}, \frac{44}{108}, \frac{3}{108}\right)$$

Interpreting the prices in terms of dollars and cents, we see the original prices were (17¢, 33¢, 50¢) and now are (56¢, 41¢, 3¢). Under **p**, we had an excess demand of commodities 1 and 2 and an excess supply of commodity 3. The transformation T has raised the prices of the first two commodities, which should lower the demand for them. T has, however, dramatically lowered the price of the third commodity. We can expect that under $T(\mathbf{p})$, the demand for commodity 3 will increase. It is certainly possible that could be more demand for commodity 3 under the price vector $T(\mathbf{p})$ than there is supply. There is no guarantee then that $T(\mathbf{p})$ will be an equilibrium set of prices.

Let's examine some of the important properties of this transformation T.

First, observe that T is a function from the set $\prod$ of normalized prices to $\prod$.

Second, consider what happens if we begin with an equilibrium price vector **p**. Then $F_h(\mathbf{p}) \leq 0$ for all h. Hence, $\delta_h(\mathbf{p}) = 0$ for all h and so $\delta(\mathbf{p}) = 0$ and $\mathbf{p} + \delta(\mathbf{p}) = \mathbf{p}$. In this case, $a = \mathbf{u} \bullet (\mathbf{p} + \delta(\mathbf{p})) = \mathbf{u} \bullet \mathbf{p} = 1$ so $T(\mathbf{p}) = \mathbf{p}$. The transformation T does not change **p**.

Third, and perhaps the most critical property is that if $T(\mathbf{p}) = \mathbf{p}$ for some price vector **p**, then **p** must be an equilibrium price vector. Here is a proof of this claim: Suppose $T(\mathbf{p}) = \mathbf{p}$ so that $\mathbf{p} = (1/a)(\mathbf{p} + \delta(p))$. Thus, $a\mathbf{p} = \mathbf{p} + \delta(\mathbf{p})$. Rewrite this equation as $(a-1)\mathbf{p} = \delta(\mathbf{p})$ and take the dot product of each side with $F(\mathbf{p})$:

$$(a-1)\mathbf{p} \bullet F(\mathbf{p}) = \delta(\mathbf{p}) \bullet F(\mathbf{p})$$

Applying Walras's Law once more, we see that $(a-1)\mathbf{p} \bullet F(\mathbf{p}) = (a-1)0 = 0$ so $\delta(\mathbf{p}) \bullet F(\mathbf{p}) = 0$. We examine this last dot product more carefully:

$$0 = \delta(\mathbf{p}) \bullet F(\mathbf{p}) = \sum_{h=1}^{l} \delta_h(\mathbf{p}) F_h(\mathbf{p})$$

but note that $\delta_h(\mathbf{p})F_h(\mathbf{p})$ is either 0 if $F_h(\mathbf{p}) < 0$ or $F_h(\mathbf{p})F_h(\mathbf{p}) = [F_h(\mathbf{p})]^2$ if $F_h(\mathbf{p}) > 0$. Thus, each term in the sum $\sum_{h=1}^{l} \delta_h(\mathbf{p})F_h(\mathbf{p})$ is nonnegative. Since the sum is 0, it must be that each term $\delta_h(\mathbf{p})F_h(\mathbf{p})$ is 0. If any $F_h(\mathbf{p}) > 0$, then $\delta_h(\mathbf{p})F_h(\mathbf{p}) = [F_h(\mathbf{p})]^2 > 0$. Hence, each $F_h(\mathbf{p})$ must be less than or equal to 0, which implies that **p** is an equilibrium price vector.

To summarize our last two results, we have established the following core theorem:

THEOREM 9.4 p is an equilibrium price vector if and only if **p** is a fixed point of T.

The centrality of *fixed points* to our question about the existence of equilibrium prices suggests this concept bears further investigation.

C. Fixed Points

DEFINITION: If f is a function from a set S to S, then an element a of S is called a *fixed point* of f if $f(a) = a$.

We illustrate the definition with a number of examples:

1. Let f be the function from the real numbers to the real numbers given by $f(x) = x^3 - 6$. The number 2 is a fixed-point of this function, since $f(2) = 2^3 - 6 = 8 - 6 = 2$.
2. The squaring function, $f(x) = x^2$, that maps the unit interval [0, 1] into itself has two fixed-points, 0 and 1, since $f(0) = 0^2 = 0$ and $f(1) = 1^2 = 1$.
3. The function that adds 1 to a number, $f(x) = x + 1$, maps the real numbers to the real numbers but has no fixed-point.
4. The identity function, $f(x) = x$, on any set S has every point of S as a fixed-point.
5. The function that rotates the unit *disk* $D = \{(x, y) : x^2 + y^2 \leq 1\}$ in the plane counterclockwise through an angle of $\pi/3$ has exactly one fixed-point, the origin $O = (0, 0)$.
6. The function that rotates the unit *circle* $D = \{(x, y) : x^2 + y^2 = 1\}$ in the plane counterclockwise through an angle of $\pi/3$ has no fixed-points.
7. Let T be the function defined on the set $\prod$ of normalized prices in a two-commodity economy, which is given by $T(x, y) = \left(\frac{3x + 6y}{10}, \frac{7x + 4y}{10}\right)$. The vector $\left(\frac{6}{13}, \frac{7}{13}\right)$ is a fixed-point for T since $T\left(\left(\frac{6}{13}, \frac{7}{13}\right)\right) = \left(\frac{\frac{18+42}{13}}{10}, \frac{\frac{42+28}{13}}{10}\right) = \left(\frac{60}{13 \times 10}, \frac{70}{13 \times 10}\right) = \left(\frac{6}{13}, \frac{7}{13}\right)$.
8. It is not difficult to see that the function T defined by $T(x, y, z) = \left(\frac{y+2}{x+y+z+9}, \frac{z+3}{x+y+z+9}, \frac{x+4}{x+y+z+9}\right)$ maps the set $\prod$ of normalized prices in a three-commodity economy into itself and has a fixed-point $\left(\frac{26}{111}, \frac{38}{111}, \frac{47}{111}\right)$.

Sometimes it is possible to show that a particular function has a fixed-point without necessarily being able to compute one. Our next example shows such an instance.

Example

Consider the continuous function on the real numbers given by $f(x) = x^{101} + 2x - 1$. Algebraically attempting to find a fixed point would mean solving the equation $x^{101} + 2x - 1 = x$, which is equivalent to $x^{101} + x - 1 = 0$. We don't have effective tools for solving a 101st-degree polynomial equation explicitly. However, the function g given by $g(x) = f(x) - x = x^{101} + x - 1$ is a continuous function with the property that $g(0) = -1$ and $g(1) = 1$ so by the Intermediate Value Theorem of calculus, there exists an number x^* between 0 and 1 with $g(x^*) = 0$ and hence, $f(x^*) = x^*$.

As another example where the geometry indicates the existence of a fixed point, consider the functions $f(x) = \cos x$ and $f(x) = x$ on the interval $[0, \pi/2]$. The graphs of the functions intersect somewhere over the subinterval $[3\pi/16, \pi/4]$. At the point of intersection (x^*, x^*) we have $\cos x^* = x^*$ so x^* is a fixed point for the cosine function. See Fig. 9.12.

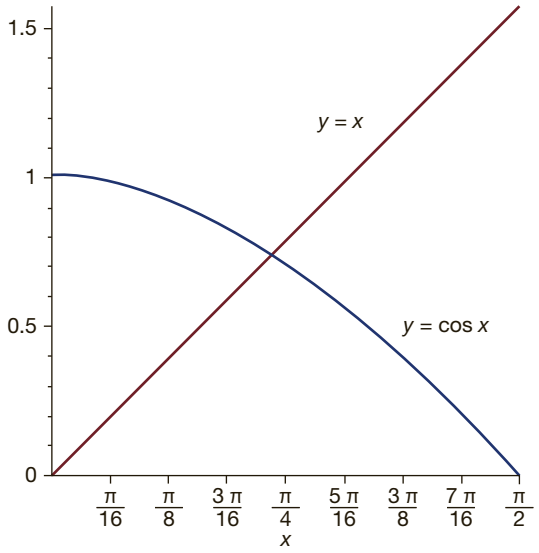


FIGURE 9.12 The cosine has a fixed point where its graph crosses the graph of the identity function.

D. Brouwer's Fixed-Point Theorem

A set S has the *fixed-point property* if every continuous function f from S into S has at least one fixed-point. Theorem 9.5 shows that the closed interval [0, 1] has the fixed-point property. Example 3 above $(f(x) = x + 1)$ demonstrates that the real line does not have the fixed-point property.

THEOREM 9.5 If f is a continuous function from the unit interval $I = [0, 1]$ of real numbers into I, then f has at least one fixed-point. Thus, the unit interval [0, 1] has the fixed-point property.

Before we examine a formal proof, let's look at an intuitive graphical argument. Imagine there is a function f with $f(0) > 0$ and $0 < f(1) < 1$ as in Fig. 9.13 where we show the points $A = (0, f(0))$ and $B = (1, f(1))$. The graph starts above the line L with equation $y = x$ and ends below this line. If f is continuous, its graph has no holes and displays no jumps. If you draw such a graph from A to B, you will have to hit the line L at least once. (Try it!) Any such intersection of L and the graph of f is a fixed-point for f.

We turn now to a more formal argument.

Proof of Theorem 9.5 Note first that $f(0) \geq 0$ and $f(1) \leq 1$. Define a new function g on I by $g(x) = f(x) - x$. Then, being the difference of two continuous functions, g is continuous with $g(0) = f(0) - 0 \geq 0$ and $g(1) = f(1) - 1 \leq 0$. By the Intermediate Value Theorem, there is a number x^* in [0, 1] with $g(x^*) = 0$. But $g(x^*) = 0$ means $f(x^*) - x^* = 0$ or, equivalently, $f(x^*) = x^*$—that is, x^* is a fixed-point for f. ◇

We can use this theorem about the fixed-point property of [0, 1] to show that in the two-commodity economy the set $\prod$ of normalized prices also has the fixed-point

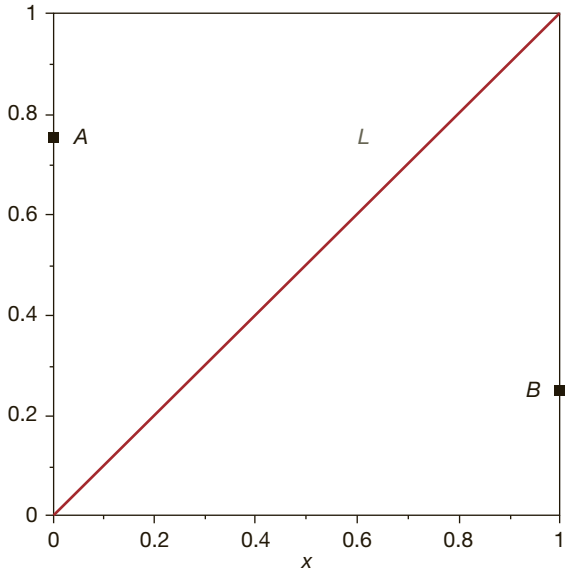


FIGURE 9.13 The graph of a continuous function f on $[0, 1]$, which contains points above and below the line $y = x$, must intersect the line in a fixed point for f.

property. To accomplish this end, suppose f is an arbitrary continuous function from $\prod$ to $\prod$. We define two auxiliary functions h from I to $\prod$ and j from $\prod$ to I by the formulas $h(x) = (x, 1 - x)$ and $j(x, y) = x$. Observe that $h \circ j$ is the identity. Note that h and j are continuous functions so that the composition $j \circ f \circ h$ is a continuous function from I to I. Since I has the fixed-point property, there is a number x^* between 0 and 1 with $(j \circ f \circ h)(x^*) = j(f(h(x^*))) = x^*$. Applying h to both sides and using the fact that $h \circ j$ is the identity function on $\prod$, we have $f(h(x^*)) = h \circ j(f(h(x^*))) = h(j()f(h(x^*))) = h(x^*)$ and thus, $h(x^*)$ is a fixed-point of f.

In 1920, the Dutch mathematician L. E. J. Brouwer (1881−1966) proved a very powerful generalization of Theorem 9.5 that implies that the set $\prod$ of normalized prices for an economy with any number of commodities has the fixed-point property. His result, known as the *Brouwer Fixed-Point Theorem*, established a category of sets in all dimensions that possess the fixed-point property. These sets also include circular disks, filled-in squares or rectangles, solid cylinders, balls, and their analogues in higher dimensional spaces. (See Exercises 27−31 for more details.)

Brouwer's result has some surprising consequences that may cause you to doubt the theorem's validity. Take two sheets with maps of the same state or country, one lying directly above the other. Crumple, without tearing, the top map and drop it onto other map. The fixed-point theorem asserts there must be at least one spot on the top map lying directly over the same location on the bottom map.

Going up a dimension, grab a cup of coffee and stir it around. When the stirring is over, Brouwer's Theorem claims that there must be some point in the coffee (not necessarily on the surface), which is in the very same spot it was before you started. If you attempt to slosh that point out of its original position, then Brouwer says you will move some other point back into the spot when it began.

The proof of Brouwer's Theorem is a little too long and complex for us to include here, but we will show an equivalent result that is a little easier to believe, the *No Retraction Theorem*.

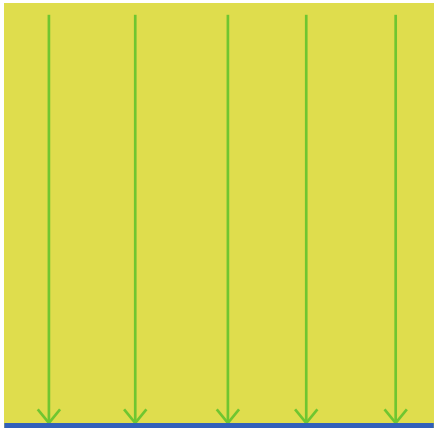


FIGURE 9.14 Retracting a solid square onto its bottom edge.

There are references to proofs of Brouwer's Theorem listed in the references. Perhaps the most accessible one, for students who have studied vector calculus, can be found in Joel Franklin's *Methods of Mathematical Economics*.

DEFINITION: If A is a set and B is a subset of A, a *retraction* of A onto B is a continuous function $f: A \to B$ such that every point of B is fixed—that is $f(b) = b$ for all points in B.

Example

Let A be the solid square $\{(a, b) : a, b \text{ are in } [0, 1]\}$ and B the bottom edge $\{(a, 0) : a \text{ is in } [0, 1]\}$. Then $f(x, y) = (x, 0)$ is a retraction of S onto R. The function f is the projection of square onto its bottom edge.

Thus, it is possible to retract a solid square onto one of its edges; see Figure 9.14. On the other hand, it is not possible to retract a disk onto its circular boundary. This result, called the *No Retraction Theorem*, is not easy to prove but it is intuitively possible. If there were such a retraction, we could continuously pull the head of drum onto its rim. This seems impossible without tearing the drum head at some point which would introduce a discontinuity. We will assume that the No Retraction Theorem is true and use it to prove the Brouwer Fixed-Point Theorem.

THEOREM 9.6 (BROUWER FIXED-POINT THEOREM FOR NORMALIZED PRICES) In any dimension, the set $\prod$ of normalized prices has the fixed-point property.

In Exercises 29 and 30, you will demonstrate that $\prod$ has the fixed-point property if and only if the unit disk has the fixed-point property. More precisely, the unit disk D in n-dimensional space $D1$ is the set of all vectors $x = (x_1, x_2, \ldots, x_n)$ such that $x_1^2 + x_2^2 + \cdots + x_n^2 \leq 1$. For $n = 1$, the unit disk is the line segment $[-1, 1]$. For $n = 2$, the unit disk is all the points on or inside the unit circle in the plane. For $n = 3$, the unit is a solid ball of radius 1. The unit disk is the set of all points within 1 unit of the origin. By the *unit sphere*, we will mean the set of all points precisely 1 unit from the origin.

Exercise 30 shows that the unit disk in n-dimensional space has the fixed-point property if and only if the set $\prod$ of normalized prices in an economy with $n+1$ commodities has the fixed-point property. It is a bit easier to work with the unit disk. We can state the No Retraction Theorem in the following language:

THEOREM 9.7 (NO RETRACTION THEOREM) There is no retraction of the unit disk onto the unit sphere.

We will now show that the No Retraction Theorem implies Brouwer's Fixed-Point Theorem. We need to show that if the No Retraction Theorem is true, then the Bouwer Fixed-Point Theorem must be true. We will actually proceed with a proof by contradiction. We begin by assuming that there is some function from D into D with no fixed-point. We will then use that function to build a retraction from D onto its boundary sphere C, contradicting the No Retraction Theorem.

Suppose then that $f : D \to D$ is a continuous function with no fixed-points. Then for each x in D, $f(x) \neq x$, so x and $f(x)$ are distinct points of D. We can then trace out a path along the line segment beginning at $f(x)$ and ending at x. Continue this line segment until it hits C. Call the point of intersection of C and this line $g(x)$. Thus, we have defined a function g from D into C. Fig. 9.15 illustrates the function g for the two-dimensional unit disk.

There are four qualitative possibilities depending on whether or not (a) x is an interior point of D or x is belongs to C and (b) $f(x)$ is an interior point of D or $f(x)$ belongs to C. Fig. 9.16 illustrates the possibilities. Note that if x belongs to C, then $g(x) = x$ whether $f(x)$ lies in the interior of D or $f(x)$ belongs to C. Thus, g is the identity on C.

Finally, note that g is the composition of the continuous function f followed by a straight line motion, which is also continuous. Hence, g is continuous and provides a retraction of D onto C, contradicting the No Retraction Theorem.

The No Retraction Theorem and the Brouwer Fixed-Point Theorem are in fact equivalent to each other. We have just shown that the No Retraction Theorem implies Brouwer's result. We now outline an argument that Brouwer Fixed-Point Theorem implies the No Retraction Theorem. Again, we will use an argument by contradiction. Suppose the No Retraction Theorem is false and let g be a retraction of D onto C. Let r be a slight rotation of C about the origin so r has no fixed-points. Then the

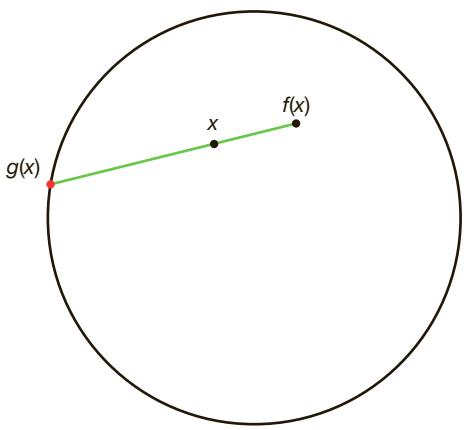


FIGURE 9.15 The function g which maps an interior point of the disk to the boundary.

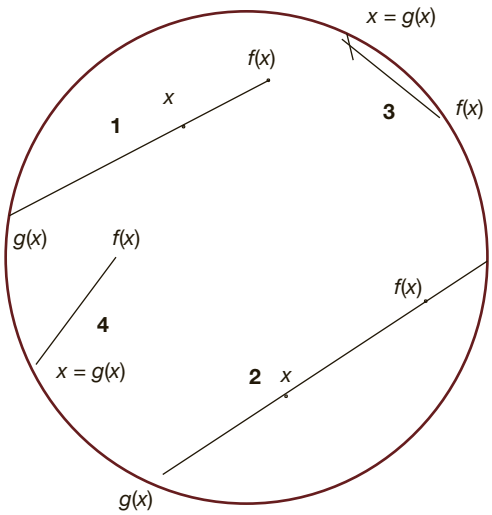


FIGURE 9.16 There are 4 possible relations between x and $f(x)$: (1) x and $f(x)$ are both in the interior of the disk; (2) x is in the interior, but $f(x)$ is on the boundary; (3) Both x and $f(x)$ are on the boundary, and (4) x is on the boundary, but $f(x)$ is in the interior.

composition $f = r \circ g$ is a fixed-point free continuous function from D into D, contradicting Brouwer's Theorem.

E. Existence of Price Equilibrium

We have seen that $\mathbf{p}$ is an equilibrium price vector if and only if $T(\mathbf{p}) = \mathbf{p}$ where T is the particular price modifying formula introduced earlier. In other words, $\mathbf{p}$ is an equilibrium price vector exactly when $\mathbf{p}$ is a *fixed-point* for the function T.

Recall that one of our basic assumptions about consumers was that their demand functions are continuous functions of prices and endowments. We obtain the total demand by adding up the individual demands of all our consumers, so the total demand is the sum of continuous functions and hence is continuous. The excess demand function is the difference of the continuous total demand and the constant vector of endowments. Since constant functions are also continuous, we find that the excess demand function F is also continuous. Each component function F_h is also continuous.

It is easy to show that if g is any continuous real valued function, then the maximum of 0 and g is also continuous. Hence, $\delta_h(\mathbf{p}) = \text{maximum}(0, F_h(\mathbf{p}))$ is continuous for each h as is the operation sending $\mathbf{p}_h$ to $\mathbf{p}_h + \delta_h(\mathbf{p})$. Therefore, $S(\mathbf{p}) = \mathbf{p} + \delta(\mathbf{p})$ is continuous. It is a straightforward argument to show then that $\mathbf{u} \bullet [\mathbf{p} + \delta(\mathbf{p})]$ is also a continuous procedure.

Putting all these pieces together, we conclude that our price transformation function T is continuous. Since T is a continuous function from $\prod$ into $\prod$, Brouwer's Fixed-Point Theorem tells us that T has at least one fixed-point. Thus, under our assumptions about consumers and their demand functions, there always exists a price equilibrium, a vector $\mathbf{p}$ of prices under which supply meets or exceeds demands for every single commodity.

We can summarize our findings with the following:

> If consumers have insatiable continuous demand functions that depend on prices and their fixed endowments, then there always exists an equilibrium set of prices that ensures that total demand will not exceed available supply.

Unlike Arrow's Theorem of Chapter 6 where a plausible set of axioms proved to be inconsistent, in this model of price equilibrium, the axioms contain no contradictions and a powerful mathematical theorem shows that an appropriate set of prices must always exist.

V. Some Remaining Questions

Almost every aspect of the model we have presented has been challenged and revised by mathematicians and economists who have sought more realistic and useful models of equilibrium in a dynamic, rapidly shifting economy. The scholarly literature in this vast and expanding field often makes use of highly sophisticated mathematics well beyond the scope of this book. We hope we have presented a sufficiently interesting introduction to the topic that will whet your appetite to learn more. The References on our text's website can help get you started.

In this section, we will discuss two troublesome features of our model and indicate how they have been ameliorated.

First, our approach to showing that there are equilibrium prices in our model relied on Brouwer's Fixed-Point Theorem. Many of the proofs of Brouwer's result, such as our reduction to the No Retraction Theorem, are proofs by contradiction. The assumption that a mapping is fixed-point free leads to contradicting another well-established truth. The situation is bit awkward. We know there are these equilibrium prices, but our proof gives no indication as to how to find or compute them.

An alternative path would be to find a *constructive proof*, one that would show us that an equilibrium price vector exists by actually describing how to locate one. Fortunately, many scientists have studied the problem of producing an algorithm, a recipe so to speak, to find equilibrium prices. The pioneer work is Herbert Scarf's book *Computation of Economic Equilibria*, which first appeared in 1973. Scarf demonstrated how to build a sequence of price vectors that was guaranteed to converge to an equilibrium. A recent survey of the field with new results on the computational complexity of finding price equilibria appears in Ye Du's 2009 thesis *Essays on the Computation of Economic Equilibria and Its Applications*.

Second, let us review first our axioms about consumer behavior. We have assumed the existence of a demand *function* from the set of prices to the set of commodities so that, given the consumer's endowment **e** and a vector **p** of prices, the consumer can tell which single commodity bundle **x** is most preferred.

A more realistic assumption is that even if the choice of commodity bundles is limited to the frontier of the consumer's budget constraint set, there is still such a vast choice of affordable bundles that it would be hard to pick out a single one. More likely, there will be set of possible bundles, all of which have exactly the same appeal to a consumer—that is, the consumer will be *indifferent* among the members of this set but prefer any commodity bundle in the set to any bundle outside the set.

To deal with this possibility, mathematicians have developed an extension to the classic definition of a function where the output is a single number or vector to outputs that are collections of objects. We call this extension a *set-valued function* or a *correspondence*. If φ is a correspondence from a set S to S, then for each x in S, $\varphi(x)$ is a subset of S.

If f is an ordinary function from S to S, then f is *continuous on S* if for every x in S, whenever $\{x_n\}$ is a sequence of points in S converging to x, the sequence $\{y_n\} = \{f(x_n)\}$ converges to $y = f(x)$. There is a corresponding notion, called *upper semicontinuity*. The

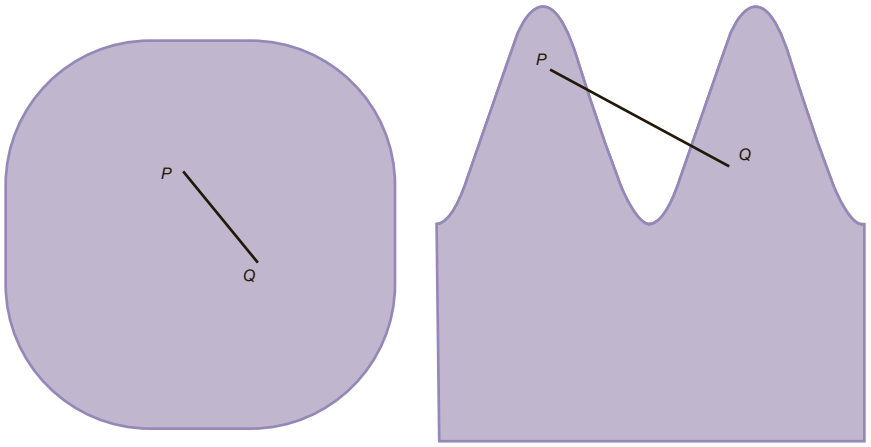


FIGURE 9.17 The shaded set on the left is convex. The shaded set on the right is not convex; there are points P and Q in that set such that there are points not in the set that belong to the line segment between P and Q.

assertion that a set valued correspondence φ is *upper semicontinuous* means that whenever $\{x_n\}$ and $\{y_n\}$ are sequences such that each y_n is an element of the set $\varphi(x_n)$ with x_n converging to x and y_n converging to y, then it must be true that y is an element of $\varphi(x)$.

There is an important extension of Brouwer's Fixed-Point Theorem that allows for a proof of the existence of price equilibrium where consumers have indifference sets of commodity bundles. To understand the statement of this generalization, we need one additional definition, the notion of a convex set in l-dimensional space.

DEFINITION: A set S is *convex* if whenever x and y are two elements in S, the entire straight line segment joining x and y lies entirely in S. Fig. 9.17 shows examples of convex and nonconvex sets.

The Kakutani Fixed-Point Theorem states that if $\varphi : S \rightarrow S$ is an upper semicontinuous correspondence where S is a non-empty, convex, closed and bounded l-dimensional set so that for every $\mathbf{x}$ in S, the set $\varphi(\mathbf{x})$ is also convex and non-empty, then φ has at least one fixed-point—that is, there is at least one $\mathbf{x}^*$ in S such that $\mathbf{x}^*$ belongs to $\varphi(\mathbf{x}^*)$.

Kenneth Arrow and Gerard Debreu [1954] used the Kakutani Theorem to establish existence of a price equilibrium in a model of a competitive economy. In 1950, John F. Nash employed the theorem in his Ph.D. thesis to prove the existence of certain types of equilibrium strategies in nonzero sum games; see Chapter 16 for more details.

VI. Historical and Biographical Notes

A. Léon Walras

Léon Walras was born December 16, 1834, in Évreux, France. His *Elements of Pure Economics* was one of the earliest comprehensive mathematical analyses of general economic equilibrium. Walras was apparently the first to examine the existence of price equilibrium as the solution of a system of simultaneous equations. Walras's argument boiled down to asserting the existence of a solution since there were more unknowns than equations. Even in the simple case of two linear equations in three unknowns, however, no solution need exist. When there are a very large collection of nonlinear equations in many

unknowns, it is certainly possible that the equations are mutually inconsistent: no assignment of values to the unknowns can satisfy all the equations.

Walras's father, the French economist Auguste Walras, encouraged his son to pursue economics with a special emphasis on mathematics. The younger Walras, like many children, ignored his father's advice at first and tried his hand at several different occupations. Initially enrolled as a student in a school of mines, he dropped out to work as a railway clerk, journalist, bank manager, and lecturer; he even published several romance novels. Walras eventually took up the study and teaching of economics where he claimed to have found "pleasures and joys like those that religion provides to the faithful." Walras retired in 1902 at age 58 from his professorship of political economy at the University of Lausanne; he died in Switzerland on January 5, 1910.

Before Walras, economists had made little attempt to show how a whole economy with many goods fits together and reaches an equilibrium. Walras believed that he could capture the essence of the problem through a system of equations whose solution would be an equilibrium. He realized, however, that a real economy might never converge to such an equilibrium. He posited an artificial market dynamic process he labeled *tâtonnement* (French for "groping") somewhat similar to our price-setting agency function T. Walras suggested a process in which a tentative price was announced and then people in the market declared how much they were willing to supply or demand at the price. If supply exceeded demand, then the price would be lowered, resulting in a greater demand and a smaller supply. Prices would "grope" toward equilibrium.

Walras also inherited his father's interest in social reform. He advocated nationalizing land, arguing that land value would always increase and that sufficient rents from that land could support the nation without taxes.

Walras spent his career as an economic thinker in Switzerland and published in French. He was far from then-important centers of economic thought in England and Germany and thus had relatively little influence during his lifetime. His work gained significant attention after his death. Walras is now considered to be one the most important of the 19th-century economists.

Public domain

Léon Walras

B. Abraham Wald

The first rigorous axiomatization of the problem and proof for the existence of price equilibria was the work of Abraham Wald, published in 1935 and considered to be one of the most distinguished achievements in mathematical economics.

Wald was born on October 31, 1902, in the city of Cluj, then part of the Austria-Hungary empire. He was home-schooled by his parents as the Hungarian school system then required attendance on Saturday, something not permitted by the Wald family's orthodox Jewish religious beliefs. The University of Vienna awarded Wald a Ph.D. in mathematics in 1931. Austrian anti-Semitism prevented Wald from securing a university position. Discrimination against Jews intensified after the Nazis took over Austria. Wald was able to emigrate to the United States in 1938; most of his close relatives, however, perished in the Holocaust.

Public domain

Abraham Wald

Wald also made important contributions to decision theory, geometry, and econometrics. He founded the field of statistical sequential analysis. During World War II, Wald made an important contribution to the problem of bomber losses to enemy antiaircraft fire. Aircraft that safely returned from missions were often damaged in similar places. Some military personnel suggested that armor should be added to those areas that displayed the most damage. Wald wisely observed that the bullet holes on the bombers that returned represented areas that were able to take damage. The bombers that had gone down must have been hit in more vulnerable spots. He concluded that extra armor should be added to those locations on the returning planes that showed the *least* damage.

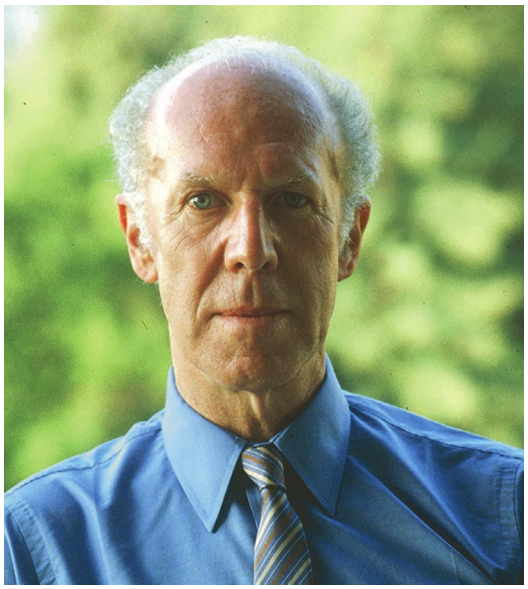
Wald and his wife died in an airplane crash on December 13, 1950, in the Nilgiri Mountains of southern India, while on an extensive lecture tour at the invitation of the Indian government.

C. Gérard Debreu

Kenneth Arrow and Gérard Debreu developed a more general model on economic equilibrium with weaker and hence more realistic assumptions about economic agents. Arrow and Debreu's 1954 paper used the Kakutani Fixed-Point Theorem to prove existence of price equilibrium for their model.

Gérard Debreu (July 4, 1921−December 31, 2004) was a French mathematician and economist who became a U.S. citizen and taught for many years at the University of California−Berkeley. Debreu received much of his university education in occupied France during World War II, enlisting in the French armed forces after D-Day. In autobiographical notes, Debreu recalled, "I had become interested in economics, an interest that was transformed into a lifetime dedication when I met with the mathematical theory of general economic equilibrium."

Two of Debreu's most important works appeared while was in his thirties: the paper with Arrow mentioned earlier and a monograph *Theory of Value*: *An Axiomatic Analysis of Economic Equilibrium*. In this short book, considered one of the most important works in mathematical economics, Debreu established an axiomatic foundation for competitive markets and used variants of the Kakutani Fixed-Point Theorem to prove the existence of equilibria.



Gérard Debreu

Debreu received the French Legion of Honor in 1976. Seven years later he was awarded the Nobel Prize in Economics (technically, The Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel) "for having incorporated new analytical methods into economic theory and for his rigorous reformulation of general equilibrium theory."

In the ceremony presenting the Nobel award to Debreu, Karl-Göran Mäler of the Royal Swedish Academy of Sciences noted:

> *You have contributed more than anyone else to our understanding of general equilibrium theory and the conditions under which there exists a general equilibrium in an abstract economy. Your insightful analysis of models of abstract economies have provided us with a general theory which may be applied to a multitude of problems offering a much broader understanding than alternative partial models could allow.*
>
> *More than anyone else, you are a symbol of a new approach to economic analysis, an approach that, while highly abstract, yields a better intuitive understanding of the basic economics. Your influence on methods, standards, and analytical techniques used by economists has been outstanding.*

"In addition to setting the agenda for General Equilibrium Theory," his colleague Robert Anderson wrote, "Gérard had a profound influence on the way *all* economic research is carried out. B. G. (Before Gérard), few economics papers had clearly specified models and virtually none showed that their models were internally consistent by exhibiting any equilibria. If you look at articles in major research journals today, virtually all clearly specify a model. . . . Gérard's insistence on mathematical clarity and rigor has had a profound effect on virtually all economic research."

In 1980, Debreu courageously undertook a potentially dangerous mission to Chile on behalf of the National Academy of Sciences to report on how scientists were being treated under the oppressive dictatorship of Augusto Pinochet, who was later tried for numerous human rights violations.

On the lighter side, Debreu volunteered to coach the Berkeley economics department football team in its first "Little Big Game" against Stanford, even though he knew nothing about American football. Playing at times in a torrential downpour, Berkeley prevailed 6−4 against the Stanford eleven, coached by Kenneth Arrow.

EXERCISES

II. A TWO-CONSUMER ECONOMY

1. Fifteen ounces of turkey and 50 ounces of meatballs are divided between Rhonda and Marc so that Rhonda initially has 8 ounces of turkey and 3 ounces of meatballs. Suppose Rhonda's indifference function has the form $\frac{a}{(1+x)^2}$, while Marc's indifference function has the form $b - \frac{x^4}{98}$.

(a) What is Marc's initial allotment of turkey and meatballs?

(b) Draw the Edgeworth box for this situation and sketch several indifference curves for Rhonda and for Marc.

(c) What are the values of a and b so the indifference curves pass through (8, 3)?

(d) Sketch the indifference curves passing through (8, 3) and shade in the bargaining space.

(e) Show that the indifference curves passing through $\left(5, \frac{750}{49}\right)$ are tangent to each other so that $\left(5, \frac{750}{49}\right)$ is a possible Pareto solution for dividing up the turkey and meatballs.

(f) Show that the function $h(x) = \frac{x^3(1+x)}{49}$ yields a contract curve.

(g) Determine at least two more Pareto solutions.

2. Julie and Brian are negotiating a redistribution of 20 ounces of cola and 50 ounces of tea. Julie currently has 5 ounces of cola and 10 ounces of soda. Julie's indifference curves have the form $y = \dfrac{a}{(1+x)^3}$, while Brian's look like $y = b - \dfrac{x^5}{100}$.

(a) What is Brian's initial allotment of cola and tea?

(b) Draw the Edgeworth box for this situation and sketch several indifference curves for Julie and Brian.

(c) What are the values of a and b so the indifference curves pass through (5, 10)?

(d) Sketch the indifference curves passing through (5, 10) and shade in the bargaining space.

(e) Show that the indifference curves passing through $\left(4, \dfrac{64}{3}\right)$ are tangent to each other so that $\left(4, \dfrac{64}{3}\right)$ is a possible Pareto solution for dividing up the cola and tea.

(f) Show that the function yields a contract curve of the form $h(x) = \dfrac{x^4(1+x)}{60}$.

(g) Determine at least two more Pareto solutions.

3. Find equation for contract curve if the indifference curves are $y = \dfrac{a}{(2+x^2)}$ and $y = b - \dfrac{x^4}{200}$.

4. Find equation for contract curve if the indifference curves are $y = \dfrac{a}{(4+x)^3}$ and $y = b - \dfrac{x^5}{200}$.

5. Suppose the indifference curves for Anne and Toby have the form $y = \dfrac{a}{(m+x)^p}$ and $y = b - \dfrac{x^q}{n}$, respectively, where a, b, m, n, p, and q are all constants greater than or equal to 1. Show that the contract curve has the equation $y = \dfrac{qx^{q-1}(m+x)}{pn}$.

6. **(a)** Indifference curves are level curves for utility functions. Suppose Anne has the utility function $u(x, y) = \ln x + 7 \ln y$ for x ounces of wine and y ounces of espresso. Show that her indifference curves have the form $y = e^{\frac{a - \ln x}{7}}$ where a is a constant.

(b) Toby's utility function if he receives x ounces of wine and y ounces of espresso is $u(x, y) = y + 8 \ln x$. If there are 15 ounces of wine and espresso available and Anne receives (x, y), then show that Toby's indifference curve has the form $y = 20 - b + 8 \ln(20 - x)$ where b is a constant.

(c) If Anne's initial holdings of wine and espresso is given by (8, 12), sketch the indifference curves passing this point and identify several points in the bargaining space.

(d) Determine the equation for the contract curve in this example.

III. AN *m*-PERSON ECONOMY

7. Find the cost of each commodity bundle $\mathbf{x}$ if the normalized price vector $\mathbf{p} = (.1, .2, .3, .4)$

(a) $\mathbf{x} = (12, 21, 6, 6)$

(b) $\mathbf{x} = (9, 13, 6, 8)$

(c) $\mathbf{x} = (8, 15, 7, 2)$

(d) $\mathbf{x} = (7, 17, 8, 2)$

8. If the prevailing prices are given by the unnormalized vector $\mathbf{p} = (3, 8, 5)$, find the wealth w of a consumer if her endowment is

(a) $\mathbf{e} = (1, 4, 7)$

(b) $\mathbf{e} = (4, 6, 2)$

(c) $\mathbf{e} = (1, 1, 9)$

(d) $\mathbf{e} = (2, 8, 3)$

9. Under consumer insatiability, a consumer would seek out the commodity bundle $\mathbf{x}$ on the boundary of the constraint set that maximizes his utility. Suppose a consumer has the utility function $u(x_1, x_2) = 3x^2y$, and endowment $\mathbf{e} = (10, 20)$ while the normalized prices are $\mathbf{p} = (.7, .3)$. Determine this consumer's wealth, budget constraint set, and most desired commodity bundle.

10. If normalized prices are $\mathbf{p} = (.8, .2)$ and a consumer's endowment is $\mathbf{e} = (4, 6)$, determine her wealth and budget constraint set. Assuming consumer insatiability and a utility function of the form $u(x, y) = (x+1)^2y^3$, find her most desired commodity bundle.

11. Some economists prefer to think of the demand function as a function of prices $\mathbf{p}$ and wealth w rather than prices and endowment $\mathbf{e}$. Show that these two views are equivalent.

12. Suppose we regard demand as a function f of prices $\mathbf{p}$ and wealth w. Then for $l = 3$, we can write $f(\mathbf{p}, w)$ as

$f(\mathbf{p}, w) = (f_1(\mathbf{p}, w), f_2(\mathbf{p}, w), f_3(\mathbf{p}, w))$ where $f_h(\mathbf{p}, w)$ is the demand for commodity h. Imagine a consumer whose demand function has the form

$$f_1(\mathbf{p}, w) = \frac{p_1}{p_1 + p_2 + p_3}\frac{w}{p_1},\ f_2(\mathbf{p}, w) = \frac{p_2}{p_1 + p_2 + p_3}\frac{w}{p_2},$$
$$f_3(\mathbf{p}, w) = \frac{p_3}{p_1 + p_2 + p_3}\frac{w}{p_3}$$

where all three prices are positive but do not necessarily sum to 1. Is the demand function homogeneous? Does it satisfy consumer insatiability?

13. Suppose we change the demand function in Exercise 12 so that

$$f_3(\mathbf{p}, w) = \frac{kp_1}{p_1 + p_2 + p_3}\frac{w}{p_3}$$

for some constant k, but retain the same functions for f_1 and f_2. Does this demand function remain homogenous? For what values of k does it satisfy consumer insatiability?

14. **(a)** Show that the sum of components in any price vector $\mathbf{q} = (q_1, \ldots, q_h, ..q_l)$ can be computed as $\mathbf{q} \bullet \mathbf{u}$ where $\mathbf{u} = (1, 1, \ldots, 1)$ is a vector, each of whose l components equals 1.

(b) Show that normalized version of $\mathbf{q}$ can be written as $\frac{\mathbf{q}}{\mathbf{q} \bullet \mathbf{u}}$.

15. Show that if $l = 3$, then $\prod$ is all the points on or inside equilateral triangle in 3-space with vertices (1, 0, 0), (0, 1, 0), and (0, 0, 1).

16. Provide a geometric description of the set $\prod$ of normalized prices in a four-commodity economy. What is the dimension of $\prod$?

17. Use Walras's Law to show that under our assumptions, if $\mathbf{p}$ is a normalized price vector and there is a commodity such that $\mathbf{p}_h = 1$, then the demand for this commodity will exactly equal the supply.

18. Suppose we have a price vector $\mathbf{p}$ where each price $\mathbf{p}_h$ is positive. If, under these prices, supply equals demand for all but one commodity, show that Walras's Law implies supply equals demand for that final commodity.

IV. EXISTENCE OF ECONOMIC EQUILIBRIUM

19. In the two-commodity case, show that if $\mathbf{r}$ is not an equilibrium price vector, then $F_1(\mathbf{r}) > 0$.

20. Complete the proof in the two-commodity case when $F_1(\mathbf{p}^*) < 0$.

21. Suppose F is a function that assigns to each k-dimensional vector $\mathbf{v}$ an l-dimensional vector, $F(\mathbf{v})$. What is the definition of continuity of F in each of the follow cases:

(a) $k = l = 1$

(b) $k = 1$ and $l > 1$

(c) $k > 1$ and $l = 1$

(d) $k > 1$ and $l > 1$

22. Suppose F is a function that assigns to each k-dimensional vector $\mathbf{v}$ an l-dimensional vector, $F(\mathbf{v})$

(a) Show that $F(\mathbf{v})$ can be written as $F(\mathbf{v}) = (f_1(\mathbf{v}), f_2(\mathbf{v}), \ldots, f_h(\mathbf{v}), \ldots, f_l(\mathbf{v}))$ where each component function $f_h(\mathbf{v})$ is a real-valued function of k variables.

(b) Prove that F is continuous if and only if each f_h is continuous.

23. Prove that if g is any continuous real valued function, then the maximum of 0 and g is also a continuous function

24. For a two-commodity economy, verify that each of the following functions is a mapping from $\prod$, the set of normalized price vectors to $\prod$ and determine at least one fixed-point for each:

(a) $f(x, y) = \left(\frac{x}{x+2}, \frac{2}{x+2}\right)$

(b) $g(x, y) = \left(\frac{2}{x+2}, \frac{x}{+x+2}\right)$

(c) $h(x, y) = \left(\frac{1}{x+1}, \frac{x}{x+1}\right)$

(d) $k(x, y) = (y, x)$

(e) $m(x, y) = \left(\frac{x+y}{1+x}, \frac{1-y}{1+x}\right)$

25. Verify that the function given by $T(x, y, z) = \left(\frac{y+2}{x+y+z+9}, \frac{z+3}{x+y+z+9}, \frac{x+4}{x+y+z+9}\right)$ does indeed map $\prod$ into $\prod$ and has $\left(\frac{26}{111}, \frac{38}{111}, \frac{47}{111}\right)$ as a fixed-point.

26. Use the Intermediate Value Theorem of elementary calculus to prove that any continuous function from the closed interval $[a, b]$ to $[a, b]$ has a fixed-point.

27. Show that the open interval (0, 1) of real numbers does not have the fixed-point property.

28. Let I be the unit interval $[0, 1]$ and B the set of points in the plane of the form (t, e^t) for $0 \le t \le 1$.

(a) Let $h : I \to B$ by $h(t) = (t, e^t)$. Show that h is continuous and each point in B is the unique image of a single point in I. [Hint: use the fact that the exponential function is strictly increasing.]

(b) Show that if (p, q) is a point on B, then $p = \ln q$.

(c) Let $j : B \to I$ by $j(p, q) = p$. Show that j is continuous and each point of I is the unique image of a point in B.

(d) Verify that the composition $h \circ j$ is the identity function on B and $j \circ h$ is the identity function on I.

(e) Let $f : B \to B$ be any continuous function from B into B. Define a new function $g : I \to I$ by $g(t) = (j \circ f \circ h)(t)$ and show that g is continuous.

(f) Why do we know that g necessarily has at least one fixed-point?

(g) Let x^* be any fixed-point of g. Show that $h(x^*)$ is a fixed-point for f.

(h) Prove that B has the fixed-point property.

29. Two sets A and B are called *topologically equivalent* if there is a one-to-one continuous function h from A onto B such that the inverse function h^{-1} is also continuous. Prove that if A and B are topologically equivalent and A has the fixed-point property, then B must also have the fixed-point property. [Hint: use an argument similar to the one outlined in Exercise 28.]

30. The standard ball D_r of radius r in l-dimensional space is the set of all vectors within r units of the origin—that is, the vector $x = (x_1, x_2, \ldots x_l)$ belongs to D_r if and only if $\mathbf{x} \bullet \mathbf{x} \leq 1$—that is, $x_1^2 + x_2^2 + \cdots + x_l^2 \leq 1$. Show that the standard ball of radius 1 in n-dimensional space is homeomorphic to the set $\prod$ of normalized prices in an economy with $n + 1$ commodities. Brouwer's original proof essentially demonstrated that the standard ball of radius 1 in n-dimensional space has the fixed-point property. Thus, this exercise and Exercise 29, along with Brouwer's Theorem, show that $\prod$ also has the fixed-point property.

31. Let A be the solid square $\{(a, b) : a, b \text{ are in } [0, 1]\}$. Find retractions of A onto

(a) the top edge $\{(a, 1) : 0 \leq a \leq 1\}$

(b) the left-hand edge $\{(0, b) : 0 \leq b \leq 1\}$

(c) the center point $(\frac{1}{2}, \frac{1}{2})$

32. Let D be the closed disk of radius 1 centered at the origin in the plane. Find a retraction of D onto the origin (0, 0) and a retraction of D onto the disk of radius ½ centered at the origin.

33. Let D_r be the standard ball of radius r in l-dimensional space. Find a retraction of D_1 onto the origin and a retraction of D_2 onto D_1.

34. Show that the closed interval $[0, 1]$ and the open interval $(0, 1)$ of real numbers are not homeomorphic. [Hint: one of them has the fixed-point property, but not the other.]

35. Does the half-open interval $(0, 1] = \{x : 0 < x \leq 1\}$ have the fixed-point property?

SUGGESTED PROJECTS

1. Investigate Abraham Wald's first proof of the existence of price equilibria. Wald's assumptions were a bit different from our presentation, which made possible a proof that used induction on the number of commodities. You can begin with Wald's original papers or the account by John Reinhard; see citations on text website.

2. Wald assumed *The Weak Axiom of Revealed Preference* (*WARP*), one of the criteria which need to be satisfied in order to make sure that the consumer is consistent with his preferences. If a bundle of goods a is chosen over another bundle b when both are affordable, then the consumer reveals that he prefers a over b. WARP says that when preferences remain the same, there are no circumstances (budget sets) where the consumer strictly prefers b over a. By choosing a over b when both bundles are affordable, the consumer reveals that his preferences are such that he will never choose b over a, regardless of income and prices. How does *WARP* play a role in Wald's proof?

3. Work through the details of a proof of Brouwer's Fixed-Point Theorem for the n-dimensional disk. Joel Franklin's treatment (referenced below) is a good starting point. A weaker version of the theorem, which is adequate for showing that $\prod$ has the fixed-point property is the Brouwer Theorem for compact, convex sets in Euclidean spaces. Fill in the details of a proof of that result.

4. Another approach to Brouwer's Fixed-Point Theorem is through the use of Sperner's Lemma. For a presentation of Sperner's result and its consequences, see Michael Henle's *A Combinatorial Introduction to Topology*. Scarf's book is also especially useful.

5. Our model of ensuring that supply was adequate for demand was based on *prices*. Another approach, which envisions a pure trading of goods without money, may possibly be constructed around the concept of a *numeraire*: select one commodity as a base and then value every other commodity in comparison to the numeraire. For example, if $l = 4$, then a set of values with commodity 2 as the numeraire might be (2, 1, 1/3, 7). In this evaluation one unit of commodity 2 can be exchanged for 2 units of commodity 1. Three units of commodity 2 would be need to swap for 1 unit of commodity 3. A consumer's wealth would then be measured in units of commodity 2. Investigate proving that there must be some way of valuing the commodities so that total demand does not exceed total supply for all commodities.

VII. Additional Historical and Biographical Notes

More information about the lives and works of Walras, Wald, and Debreu are easily found. For Léon Walras, I advise the article on Walras in the *International Journal of Social Sciences*, William Darity, ed. (Macmillan, 2007). It can be also be found online at www.encyclopedia.com/topic/Leon_Walras.aspx. Oskar Morgenstern has a useful memoir, "Abraham Wald, 1902−1950," in *Econometrica* **19** (1951): 361−367. Debreu submitted a short autobiography at the time of his receipt of the Nobel Prize. You can find it at www.nobelprize.org/nobel_prizes/economic-sciences/laureates/1983/debreu-bio.html. There are additional personal recollections of Debreu in the Spring 2005 issue of *The Econ Exchange*: *News and Notes of the Department of Economics* (University of California−Berkeley) **8**(1), www.econ.berkeley.edu/sites/default/files/econexchangespring05.pdf.

You can find a listing of references and suggestions for additional reading on the book's website, www.wiley.com/college/olinick

Elementary Probability

CHAPTER 10

> The most important questions of life are, for the most part, really only questions of probability. Strictly speaking, one may even say that nearly all our knowledge is problematical; and in the small number of things which we are able to know with certainty, even in the mathematical sciences themselves, induction and analogy, the principal means of discovering truth, are based on probabilities, so that the entire system of human knowledge is connected with this theory.
>
> —Pierre-Simon de Laplace

I. The Need for Probability Models

The deterministic and axiomatic models developed in earlier chapters show that both types of models can serve to give concise and precise descriptions of some real-world situations. Deterministic models have an added feature of being predictive in nature, while the best that axiomatic models seem to do is guarantee the existence or uniqueness of certain kinds of sets or functions.

The usefulness of a model increases if that model gives some new information not yet observed about the situation it is supposed to represent. The predictions of the model can be tested against what actually happens in the real world. Refinements can then be made in the model and better understanding gained of the real-world problem.

The deterministic models of Chapters 1–5 are typical of the type one sees in the natural, physical, and social sciences. They consist of systems of differential equations, the mathematical tool that has been most useful in the study of physical systems. Mathematically, differential equations of the type we have examined assert that once the equations and the initial conditions are specified, then the state of the corresponding system at any later moment is completely determined.

The main criticism of the deterministic approach to the study of social problems lies precisely in this feature of the mathematics. Social problems deal with individuals or groups of individuals and we can never completely predict the exact future behavior of any person in a specific situation, no matter how well we understand the situation and the person. People are not particles, this objection concludes, and the equations of physics cannot be used to describe human actions.

There are at least two responses to such objections. First, the mathematical modeler never makes the grandiose claim that his equations *completely* describe the real-world

situation. He realizes that he has undoubtedly left out some important variables and that he has simplified the interactions among the variables he has included. At best, he hopes the qualitative nature of his quantitative results will mimic what happens in the real world.

There is a second response to the objection. Accept the premise that determinism is no fit way to describe social phenomena. This does not mean that mathematical social science is inherently any less rigorous than physics. Physicists have come, in this century, to the belief that determinism is also not possible in this most deterministic of all sciences. Nobel Laureate Richard Feynman emphasizes this realization in his *Lectures on Physics* [1965]:

> *We would like to emphasize a very important difference between classical and quantum mechanics. We have been talking about the probability that an electron will arrive in a given circumstance. We have implied that in our experimental arrangement (or even in the best possible one) it would be impossible to predict exactly what would happen. We can only predict the odds! This would mean, if it were true, that physics has given up on the problem of trying to predict exactly what will happen in a definite circumstance. Yes! Physics* has *given up. We do not know how to predict what would happen in a given circumstance, and we believe that it is impossible—that the only thing that can be predicted is the probability of different events. It must be recognized that this is a retrenchment in our earlier ideal of understanding nature. It may be a backward step, but no one has seen a way to avoid it. . . . We suspect very strongly that it is something that will be with us forever . . . that this is the way nature really is.*

This chapter begins the study of probabilistic models of human behavior by introducing the basic tools of probability. Before starting, it should be emphasized that while mathematical social scientists believe that the existence of human "free will" implies that the probabilistic approach is the more correct style of modeling, deterministic approaches will still be employed. Most social phenomena are quite complex. Accurate mathematical models must also be complex. A deterministic model based on a given set of axioms may be simpler to analyze than a probabilistic one. In addition, probabilistic models are often analyzed by approximating them with more tractable deterministic ones. A successful modeler cannot dismiss either approach. In Section III and again in Chapters 12 and 14, we will present comparisons of deterministic and probabilistic attacks on the same problems.

II. What Is Probability?

A. Fundamental Definitions

This section (parts A–F) outlines the bare minimum of probability theory on finite sample spaces. Sources for complete treatments are listed in the References at the end of the chapter.

DEFINITION Let E be a set with a finite number of elements. A *probability measure on* E is defined to be a real-valued function Pr whose domain consists of all subsets of E and that satisfies three rules:

1. $\Pr(E) = 1$
2. $\Pr(X) \geq 0$ for every subset X of E
3. $\Pr(X \cup Y) = \Pr(X) + \Pr(Y)$ for every pair of disjoint subsets X and Y of E

A finite set E together with a probability measure is called a *sample space*.

The definitions will be illustrated by several examples.

Example 1

Let E be the set of possible outcomes in an experiment consisting of flipping a coin and noting which side faces up when the coin lands. Then E has two elements, h and t, corresponding to "heads" and "tails." Note that $E = \{h, t\}$ and there are four subsets: E, $\varnothing$, $\{h\}$, $\{t\}$. An assumption that the coin is fair, that is, there is as much chance of a head as of a tail, may be reflected by the probability measure:

$$\Pr(\varnothing) = 0,\ \Pr(E) = 1,\ \Pr(\{h\}) = \Pr(\{t\}) = 1/2$$

Example 2

Suppose that the coin of Example 1 has been weighted so that heads appear twice as often as tails. Then we might assign a different probability measure:

$$\Pr(\varnothing) = 0,\ \Pr(E) = 1,\ \Pr(\{h\}) = 2/3,\ \Pr(\{t\}) = 1/3$$

Note that Examples 1 and 2 are two different sample spaces with the same underlying set.

Example 3

There is an urn with four marbles. Each marble has a different color: green, blue, red, or white. Reach your hand into the urn and, without looking, remove one marble. If E represents the set of possible outcomes of this experiment, the E consists of four elements, represented by the letters g, b, r, and w, corresponding to the color of the marble selected. Rather than list the probability measures of all 16 subsets of E, we may define the probability of any subset X by

$$\Pr(X) = \frac{\text{Number of distinct elements in } X}{4}$$

As an exercise, check whether this definition satisfies the three conditions of a probability measure.

Example 4

Replace the urn of Example 3 by one holding 4 green marbles, 3 blue ones, 2 red ones, and 1 white one. Otherwise the experiment is the same and the set E is again the collection of all possible outcomes of removing one marble. Call the outcomes g, b, r, w *elementary events* and assign them weights of .4, .3, .2, and .1, respectively. Define a probability measure on E by letting $\Pr(X)$ be equal to the sum of the weights of the distinct elementary events in X. For example, if $X = \{g, w\}$, then $\Pr(X) = .4 + .1 = .5$. Check whether this assignment of probabilities also satisfies Rules (1)—(3).

The elementary events in Example 3 were each assigned the same measure, 1/4. This is an illustration of an *equiprobable measure* that occurs whenever each of the finite number of elements of a set E has the same weight. The probability measure in the equiprobable situation has a very simple form: If E has n distinct elements and X is a subset containing r of these elements, then $\Pr(X) = r/n$. Note that the equiprobable measure was also used in Example 1, but not in Examples 2 and 4.

It is useful to list here some of the elementary laws of probability implied by the definition of a probability measure. These are gathered together in the following theorem, whose proof is left as an exercise.

THEOREM 1 If Pr is a probability measure on a finite set E, then the following statements are true for all subsets $X, Y, X_1, X_2, \ldots, X_k$ of E:

1. If $X \subseteq Y$, then $\Pr(X) \leq \Pr(Y)$.
2. $\Pr(\varnothing) = 0$.
3. $\Pr(X \cup Y) = \Pr(X) + \Pr(Y) - \Pr(X \cap Y)$.
4. $\Pr(X^C) = 1 - \Pr(X)$ where X^C is the complement of X—that is, $X^C = E - X$.
5. $\Pr(Y) = \Pr(X \cap Y) + \Pr(X^C \cap Y)$.
6. If $X_1 X_2, \ldots, X_k$ are mutually disjoint subsets of E, then

$$\begin{aligned} \Pr(X_1 \cup X_2 \cup \ldots \cup X_k) &= \Pr(X_1) + \Pr(X_2) + \cdots + \Pr(X_k) \\ &= 1 - \Pr({X_1}^C \cap {X_2}^C \cap \ldots \cap {X_k}^C) \end{aligned}$$

B. Conditional Probability

As you read this section and glance through probability textbooks, you will see many examples having to do with pulling objects out of urns. Probabilists have no particular psychological hang-ups about urns. Urns simply provide a convenient mechanism for conceptualizing and clarifying many of the important concepts of the subject. Be patient; we will soon be dealing with people and not urns.

Imagine, then, an urn containing six red marbles—numbered 1, 2, 3, 4, 5, 6 and ten green marbles, numbered from 1 to 10. Other than color, the marbles are identical in shape

and appearance. As usual, reach into the urn without looking and remove one marble. What is the probability that the selected marble is labeled with a "3"?

A reasonable answer to this question is 2/16. A reasonable explanation is, "There are 16 marbles and I am no more likely to pick one than another, so I assume I am working with the equiprobable measure. Since the set E has 16 elements, each has probability 1/16. The subset that corresponds to obtaining a marble labeled 3 has exactly two elementary events: the green marble 3 and the red marble 3. Thus, the probability is $1/16 + 1/16 = 2/16$."

Very good. Nothing new so far. Suppose, however, that you observed that the selected marble was red before you were asked "What is the probability that the selected marble bears a '3' on it?" The reasonable answer to the question is now 1/6 since there are six red marbles, each equally likely to be chosen, and exactly one of them is labeled "3."

Different answers to the same question are appropriate because different amounts of information were given in each of the situations. Additional information often changes estimates of probabilities. The concept of *conditional probability* makes this precise.

DEFINITION Let Pr be a probability measure defined on a set E. If X and Y are any two subsets of E with $\Pr(X) > 0$, then the *conditional probability of Y given X*, denoted $\Pr(Y|X)$ is defined by

$$\Pr(Y|X) = \frac{\Pr(Y \cap X)}{\Pr(X)}$$

If $\Pr(X) = 0$, then the conditional probability of Y given X is not defined.

To illustrate the definition with the example just given, let Y be the subset corresponding to "The marble is labeled 3" and X the subset corresponding to "The marble is red." Then $\Pr(Y) = 2/16$, $\Pr(X) = 6/16$, and $\Pr(Y \cap X) = 1/16$ since there is exactly one marble that is red and labeled "3." The conditional probability of Y given X is

$$P(Y|X) = \frac{Pr(Y \cap X)}{Pr(X)} = \frac{1/16}{6/16} = \frac{1}{6}$$

agreeing with the verbal explanation first given.

In this case, note that the conditional probability of X *given* Y is also defined and is equal to

$$P(X|Y) = \frac{Pr(X \cap Y)}{Pr(Y)} = \frac{1/16}{2/16} = \frac{1}{2}$$

To see that this is a reasonable result, consider that $\Pr(X|Y)$ is the answer to the question, "If you are told that the marble is labeled '3', what is the probability that it is red?"

The calculations just given illustrate the critical warning that, in general, $\Pr(Y|X) \neq \Pr(X|Y)$.

Example 5

Two weeks before the state primary to choose a single nominee for election to the U.S. Senate, there were four candidates. Political experts gave Oppenheim a .4 chance of winning, Mazzoli a .3 chance, Levine a .2 chance, and Newman a .1 chance. Just prior to the election, a grand jury indicts Levine, charging him with accepting illegal campaign contributions. If Levine withdraws from the race, how would this affect the chances of winning of the remaining three candidates?

Solution

In the absence of other information, we may assume that we have a set with four elements, Oppenheim, Mazzoli, Levine, and Newman, with weights of .4, .3, .2, and .1 measuring the probability of each winning. We will compute the chances for Oppenheim winning if Levine withdraws.

By Theorem 1, $\Pr(\text{Levine loses}) = 1 - \Pr(\text{Levine wins}) = 1 - .2 = .8$. To find the conditional probability that Oppenheim wins given that Levine loses, use the definition of conditional probability:

$$\begin{aligned}\Pr(&\text{Oppenheim wins and Levine loses})\\ &= \frac{Pr(\text{Oppenheim wins and Levine loses})}{Pr(\text{Levine loses})}\\ &= \frac{\Pr(\text{Oppenheim wins})}{\Pr(\text{Levine loses})}\\ &= \frac{.4}{.8} = \frac{1}{2}\end{aligned}$$

In this computation, we use condition (5) of Theorem 1 in the form

$$\begin{aligned}\Pr(&\text{Oppenheim wins})\\ &= \Pr(\text{Oppenheim wins and Levine wins}) + \Pr(\text{Oppenheim wins and Levine loses})\\ &= 0 + \Pr(\text{Oppenheim wins and Levine loses})\end{aligned}$$

since the subset corresponding to "Oppenheim wins and Levine wins" is empty.

C. Bayes's Theorem

The equation defining conditional probability can be rewritten as

$$(*) \quad \Pr(Y \cap X) = \Pr(Y|X)\Pr(X)$$

This equation is useful in computing $\Pr(X|Y)$ in certain instances when the probability $\Pr(Y|X)$ is given, since

$$(**) \quad \Pr(X|Y) = \frac{Pr(X \cap Y)}{Pr(Y)} = \frac{Pr(Y \cap X)}{Pr(Y)} = \frac{Pr(Y|X)Pr(X)}{Pr(Y)}$$

Example 6

A multiple-choice exam has four suggested answers to each question, only one of which is correct. A student who has done her homework is certain to identify the correct answer. If a student skips her homework, then she chooses an answer at random. Suppose that two-thirds of the class has done the homework. In grading the test, the teacher observes that Julie has the right answer to the first problem. What is the probability that Julie did the homework?

Solution

Let X denote "Julie has done the homework" and Y denote "Julie has the right answer." The information given in the problem translates into three probability statements:

$$\Pr(X) = \frac{2}{3} \quad \Pr(Y|X) = 1 \quad \Pr(Y|X^C) = \frac{1}{4}$$

The question asks for the computation of $\Pr(X|Y)$. From Theorem 1, we have $\Pr(Y) = \Pr(Y \cap X) + \Pr(Y \cap X^C)$ and two uses of the equation
$\Pr(Y \cap B) = \Pr(Y|B)\Pr(B)$, with $B = X$ and $B = X^C$ respectively
give

$$\Pr(Y \cap X) = \Pr(Y|X)\Pr(X)$$

and

$$\Pr(Y \cap X^C) = \Pr(Y|X^C)\Pr(X^C)$$

Putting this information together with the definition of conditional probability gives the answer:

$$\begin{aligned}\Pr(X|Y) &= \frac{\Pr(Y|X)\Pr(X)}{\Pr(Y|X)\Pr(X) + \Pr(Y|X^C)\Pr(X^C)} \\ &= \frac{(1)(2/3)}{(1)(2/3) + (1/4)(1/3)} = \frac{8}{9}\end{aligned}$$

The type of calculation used to solve the question of Example 6 occurs in a great many applications. The general rule that underlies it is called Bayes's Theorem and is the content of Theorem 2. It is named after the Reverend Thomas Bayes (1702–1761) and first appeared in print in 1763.

THEOREM 2 (Bayes's Theorem)
Let Pr be a probability measure defined on a set E and suppose Y is a subset of E with $\Pr(Y) > 0$. If $X_1, X_2, \ldots, X_k$ is any collection of mutually disjoint subsets of E whose union is all of E, and each X_i has positive probability, then

$$\Pr(X_j|Y) = \frac{Pr(Y|X_j)\Pr(X_j)}{\sum_{i=1}^{k} \Pr(Y|X_i)\Pr(X_i)}$$

Proof of Theorem 2 Write

$$\begin{aligned} Y &= Y \cap E = Y \cap (X_1 \cup X_2 \cup \ldots \cup X_k) \\ &= (Y \cap X_1) \cup (Y \cap X_2) \cup \ldots \cup (Y \cap X_k) \end{aligned}$$

and use Theorem 1 to obtain

$$\Pr(Y) = \sum_{i=1}^{k} \Pr(Y \cap X_i).$$

Now use (*) to write $\Pr(Y \cap X_i) = \Pr(Y|X_i)\Pr(X_i)$ so that $\Pr(Y) = \sum_{i=1}^{k} \Pr(Y|X_i)\Pr(X_i)$. An application of equation (**) completes the proof. ◇

Example 7 (Mass Screening for Drug Use)

The chief of a large metropolitan police force has evidence that 2% of his officers are heroin users. The chief wants to identify the drug users on his force so he can fire them. He orders mandatory drug testing each week for everyone on the force. The drug test correctly identifies users 95% of the time and also correctly identifies nonusers with 90% accuracy. Detective Joe Friday's test has a positive outcome. Should the chief fire Joe? What is the probability that Joe is actually a heroin user?

Solution

Given that the accuracy levels of the test are so high, many people's first response is that someone who tests positive is in all likelihood a heroin user. A careful analysis tells another story.

Let Pr(Heroin User) be the probability that a random chosen person is a heroin user, Pr(Positive Test) be the probability of a positive test—that is, the test indicates the person uses the drug. The information that we are given in the problem is Pr(Heroin User) = .02, Pr (Positive Test | Heroin User) = .95 and Pr(Negative Test | not Heroin User) = .90. We seek the conditional probability Pr(Heroin User | Positive Test).

In drug testing, the term "sensitivity" describes the likelihood of a true positive—that is that the test detects heroin in a user. Sensitivity in this case is Pr(Positive Test | Heroin User) = .95. The term "specificity" refers to the probability of a true negative—that is, the

test finds no heroin in a non-user. In our example, the specificity is .90. We would like to have numbers close to 1 for both the sensitivity and the specificity. These numbers would give small values for the possible outcomes of the test that are wrong: False Positive and False Negatives. In the case of a False Positive, an innocent person may be labeled a drug user. In the event of a False Negative, a true user escapes detection.

$$\begin{aligned}\text{Now } \Pr(\text{Positive Test}) &= \Pr(\text{Heroin User} \cap \text{Positive Test}) \\ &\quad + \Pr(\text{Not Heroin User} \cap \text{Positive Test}) \\ &= \Pr(\text{Heroin User})\Pr(\text{Positive Test}|\text{Heroin User}) \\ &\quad + \Pr(\text{Not Heroin User})\Pr(\text{Positive Test}|\text{Not Heroin User}) \\ &= (.02)(.95) + (.98)(.10) = .019 + .098 = .117\end{aligned}$$

Thus, the likelihood that Joe is a heroin user is

$$\begin{aligned}\Pr(\text{Heroin User}|\text{Positive Test}) &= \frac{\Pr(\text{Heroin User} \cap \text{Positive Test})}{\Pr(\text{Positive Test})} \\ &= \frac{.019}{.117} = \sim .162\end{aligned}$$

On the basis on the drug test alone, there is just over a 16% chance that Joe is a heroin user. The vast majority (nearly 84%) of the positive tests are false positives.

The results of this example are fairly typical of mass screening for a particular trait that occurs with low frequency in a large population. Even with high levels of sensitivity and specificity, the rate of false positives will be large.

One final example will illustrate the use of Bayes's Theorem to assess the reliability of eyewitness testimony.

Example 8

The city of Metropolis has three taxi companies, each of which uses only cabs of the color that matches the company name: Yellow Cab, Blue Cab, Green Cab. Sixty percent of the cabs are yellow, 37% are blue, and the remaining 3% are green. On a dark and stormy night a cab was involved in a hit-and-run accident. There was a single eyewitness who saw the cab fleeing the scene of the accident. He identified the cab as a green one. Tests show that people report cab color with the accuracies shown in Table 10.1:

Table 10.1 Probabilities of What Witnesses Report

Actual Color	*Says Blue*	*Says Green*	*Says Yellow*
Yellow	.1	.1	.8
Blue	.8	.15	.05
Green	.08	.8	.12

What is the probability that the cab involved in the accident was an indeed a green one, as our witness said. Is it as high as 80%, as the data in Table 10.1 indicate?

Solution

We want to find Pr(Cab was Green | Witness says Green)

By the definition of conditional probability, this probability is

$$\frac{\Pr(\text{Cab was Green and Witness says Green})}{\Pr(\text{Witness says Green})}$$

$$\begin{aligned}\Pr(\text{Witness says Green}) &= \Pr(\text{Witness says Green AND Cab was Yellow}) \\ &\quad + \Pr(\text{Witness says Green AND Cab was Blue}) \\ &\quad + \Pr(\text{Witness says Green AND Cab was Green}) \\ &= \Pr(\text{Witness says Green})|\text{Cab was Yellow})\Pr(\text{Cab was Yellow}) \\ &\quad + \Pr(\text{Witness says Green})|\text{Cab was Blue})\Pr(\text{Cab was Blue}) \\ &\quad + \Pr(\text{Witness says Green})|\text{Cab was Green})\Pr(\text{Cab was Green}) \\ &= (.1)(.6) + (.15)(.37) + (.8)(.03) = .06 + .0555 + .024 = .1395\end{aligned}$$

Thus, $\Pr(\text{Cab was Green}|\text{Witness says Green}) = \frac{.024}{.1395} = .172$. In the absence of other evidence, a jury would be more accurate at assigning only a 17% chance to the witness's being correct in his claim that the cab was green.

D. Independent Events

Knowledge about some aspects of the outcome of an experiment on a sample space can influence the estimate of the probability of other aspects of the outcome. This influence is measured using conditional probability. Sometimes, however, the extra knowledge does not influence the estimate.

Consider, as an example, an experiment consisting of flipping a coin and rolling a die. The outcome of the experiment consists of two observations: a head or a tail for the coin, and a number between 1 and 6 for the die. The coin in no way affects the die, so the answer to the question "What is the probability that the die shows a 3?" is the same whether or not you know how the coin landed. More exactly, the probability that the die shows a 3 given the coin lands heads is the same as the probability that the die shows a 3 given no information about the coin. A probabilist would say that the coin flip and die roll are *independent* of each other. The general definition looks like this:

DEFINITION Let Pr be a probability measure on a set E. If X and Y are subsets of E with $\Pr(X) > 0$ and $\Pr(Y) > 0$, then X and Y are *independent events* if $\Pr(Y|X) = \Pr(Y)$.

If X and Y are independent, then

$$\Pr(X|Y) = \frac{\Pr(Y|X)\Pr(X)}{\Pr(Y)} = \frac{\Pr(Y)\Pr(X)}{\Pr(Y)} = \Pr(X)$$

The definition also gives the very important multiplicative rule for independent events.

THEOREM 3 Suppose $\Pr(X) > 0$ and $\Pr(Y) > 0$. Then X and Y are independent if and only if $\Pr(X \cap Y) = \Pr(X)\Pr(Y)$.

Proof of Theorem 3 Recalling equation (*) of Section II.C, we have

$$\Pr(X \cap Y) = \Pr(Y \cap X) = \Pr(Y|X)\Pr(X)$$

so that $\Pr(X \cap Y) = \Pr(Y)\Pr(X)$ if and only if $\Pr(Y|X) = \Pr(Y)$. ◇

Since the definition of independent events makes use of conditional probabilities, it must be restricted to events with positive probabilities. However, the equation $\Pr(X \cap Y) = \Pr(X)\Pr(Y)$ may hold true for events with zero probabilities. For this reason, many probability theorists *define* two events X and Y to be independent if the multiplicative rule is valid.

There is a standard mistake that many students make in thinking about independence. The independence of two events is not determined strictly from the intrinsic nature of the events. Independence is also a function of the probability measure that has been assigned to the original set of outcomes. Two events may be independent under one probability measure, but not independent under another measure. Consider the next two examples.

Example 9

A pyramid is a solid figure with four triangular faces. Suppose the faces are labeled with the letters a, b, c, d. Roll the pyramid and observe which triangle faces the ground when the pyramid comes to rest. The set E of outcomes may be denoted by $E = \{a, b, c, d\}$. Let X be the subset $\{a, c\}$ and Y the subset $\{b, c\}$. The $X \cap Y = \{c\}$. If Pr is the equiprobable measure on E, then $\Pr(X \cap Y) = 1/4$ while $\Pr(X)\Pr(Y) = (2/4)(2/4) = 1/4$. Thus, X and Y are independent events in this sample space.

Example 10

Consider the same situation as Example 9, except that the probability measure is defined by assigning a, b, c, d weights of .4, .4, .1, .1, respectively. Then $\Pr(X \cap Y) = .1$ while $\Pr(X)\Pr(Y) = (.5)(.5) = .25$. Thus, X and Y are not independent in this sample space.

By making use of the multiplicative rule, the concept of independence is easily extended to more than two events. Three events X, Y, Z will be called *mutually independent* if each pair of events is independent and

$$\Pr(X \cap Y \cap Z) = \Pr(X)\Pr(Y)\Pr(Z)$$

More generally, a set of events $X_1, X_2, \ldots, X_n$ in a sample space is *mutually independent* if the probability of the intersection of any k distinct events in the set is equal to the product of the probabilities of the events where $k = 2, 3, \ldots, n$.

Independence is an important idea in discussion of situations in which the same experiment is repeated under identical conditions a number of times. Suppose, for example, that a fair coin is tossed three times. It is reasonable to assume that successive tosses of the coin do not influence each other: the coin has no memory of how it has landed before on earlier tosses. In other words, the sequence of outcomes is a mutually independent set. Let H_i be the subset corresponding to obtaining a head on the ith toss, for $i = 1, 2, 3$, then the probability of obtaining heads on all three tosses is $\Pr(H_1 \cap H_2 \cap H_3)$. By the assumption of independence this is equal to $\Pr(H_1)\Pr(H_2)\Pr(H_3) = \frac{1}{2} \bullet \frac{1}{2} \bullet \frac{1}{2} = \frac{1}{8}$.

Modifying the example, suppose the coin has been weighted so the probability of a head on a single toss is $\frac{1}{3}$. If the coin is tossed three times, what is the probability that it will land heads exactly once? If A is the subset corresponding to obtaining exactly one head in three tosses, then A can be written as the union of three mutually disjoint subsets

$$A = (H_1 \cap T_2 \cap T_3) \cup (T_1 \cap H_2 \cap T_3) \cup (T_1)T_2 \cap H_3)$$

where T_i indicates a tail on toss i. By condition (6) of Theorem 1 and the assumption of independence, we have

$$\begin{aligned} \Pr(A) &= \Pr(H_1)\Pr(T_2)\Pr(T_3) + \Pr(T_1)\Pr(H_2)\Pr(T_3) + \Pr(T_1)\Pr(T_2)\Pr(H_3) \\ &= \left(\frac{1}{3}\right)\left(\frac{2}{3}\right)\left(\frac{2}{3}\right) + \left(\frac{2}{3}\right)\left(\frac{1}{3}\right)\left(\frac{2}{3}\right) + \left(\frac{2}{3}\right)\left(\frac{2}{3}\right)\left(\frac{1}{3}\right) = \frac{12}{27} \end{aligned}$$

As a significant generalization of this example, consider an experiment with precisely two outcomes with associated probabilities p and q, where p and q are nonnegative numbers with $p + q = 1$. Call the outcome with probability p a "success" and the other outcome a "failure." Repeat this experiment a number of times in such a manner that the outcomes of any one experiment in no way affect the outcomes in any other experiment—that is, assume the sequence of outcomes forms a mutually independent set. Let X_i represent the outcome of a success on the ith trial of the experiment and Y_i the outcome of a failure on the ith trial. Then $\Pr(X_i) = p$ and $\Pr(Y_i) = q = 1 - p$ for each i.

Suppose the experiment is repeated four times. The probability that there are successes on the first and fourth trials and failures on the second and third is given by

$$\Pr(X_1 \cap Y_2 \cap Y_3 \cap X_4)$$

which by independence is equal to

$$\Pr(X_1)\Pr(Y_2)\Pr(Y_3)\Pr(X_4) = pqqp = p^2q^2 = p^2(1 - p)^2$$

It should be clear that any other prescribed sequence of two successes and two failures in four trials will also have probability $p^2(1 - p)^2$.

In general, if the experiment is repeated n times, then the probability of obtaining a prescribed sequence of exactly k successes and $n - k$ failures will be $p^k q^{n-k} = p^k(1 - p)^{n-k}$.

A related question concerns the probability of obtaining exactly k successes in n trials. This probability will be $p^k q^{n-k}$ multiplied by the number of distinct ways one can prescribe a sequence of k successes and $n-k$ failures. This number is equal to

$$\frac{n!}{k!(n-k)!} = \frac{n(n-1)(n-2)\cdots(n-k+1)(n-k)(n-k-1)\cdots(3)(2)(1)}{k(k-1)(k-2)\cdots 1(n-k)(n-k-1)\cdots 1}$$
$$= \frac{n(n-1)(n-2)\cdots(n-k+1)}{k(k-1)(k-2)\cdots 1}$$

(See Exercises 25–27 for its determination.) Thus, the number of ways of exactly obtaining 3 successes in 7 trials is computed by letting $n=7$ and $k=3$ so that $n-k+1=5$. The number of ways is then $\frac{(7)(6)(5)}{(3)(2)(1)} = 35$. The probability that a fair coin will give 3 heads and 4 tails in 7 tosses is then $35\left(\frac{1}{2}\right)^3\left(\frac{1}{2}\right)^4 = \frac{35}{128}$.

E. Expected Value

The discussion of gambles in the development of utility theory (Chapter 8) presented intuitively an idea of "expected value" or "expectation" of a gamble. This was a number meant to measure the average result of the gamble if it is made many times. In this section we formally extend this concept to more general probabilistic situations.

DEFINITION Let Pr be a probability measure on a finite set E. A *random variable* is a real-valued function R defined on E. Let $a_1, a_2, \ldots, a_k$ be the finite set of distinct values taken on by the function R. Then the *expected value* or *expectation* of R, denoted $EV(R)$, is the number

$$EV(R) = \sum_{i=1}^{k} a_i Pr(R = a_i) = a_1 Pr(R = a_1) + \cdots + a_k Pr(R = a_k)$$

The mysteries of this equation will disappear after considering the next few examples.

Example 11

Roll a fair die and let R be equal to the number showing on the top of the die when it comes to rest. Then R takes on the values 1, 2, 3, 4, 5, and 6. The event "$R=3$" is just the event that the die shows a 3 and thus has probability 1/6, so that $\Pr(R=3) = 1/6$. Similarly, R takes on each of the other values with probability 1/6. The expected value of R is given by

$$EV(R) = 1\left(\frac{1}{6}\right) + 2\left(\frac{1}{6}\right) + 3\left(\frac{1}{6}\right) + 4\left(\frac{1}{6}\right) + 5\left(\frac{1}{6}\right) + 6\left(\frac{1}{6}\right) = \frac{21}{6} = \frac{7}{2}$$

Example 12

Roll the die of Example 10, but this time let R be the square of the number that appears on top. The function R takes on the values 1, 4, 9, 16, 25, and 36, each with probability 1/6. The expected value of this random variable is

$$1\left(\frac{1}{6}\right)+4\left(\frac{1}{6}\right)+9\left(\frac{1}{6}\right)+16\left(\frac{1}{6}\right)+25\left(\frac{1}{6}\right)+36\left(\frac{1}{6}\right)=\frac{91}{6}$$

Example 13

Suppose you win \$3 every time the fair die shows an odd number and lose \$2 each time an even number appears. The set E of outcomes of rolling the die is the same, $E=\{1, 2, 3, 4, 5, 6\}$. Define the random variable R on E by $R(1)=R(3)=R(5)=+3$, and $R(2)=R(4)=R(6)=-2$. Then $\Pr(R=3)=\Pr(\{1, 3, 5\})=1/6+1/6+1/6=1/2$, and $\Pr(R=-2)$ is also 1/2. Then the expected value of R is $3(1/2)+(-2)(1/2)=1/2$. The interpretation of expected value here is that if you roll the die a great many times, you can expect to win, on average, 50¢ on each roll.

Examples 11–13 show that the expected value of a random variable need not be one of the values the random variable actually takes on.

Example 14

Brian's roommate accidentally knocks Brian's passport into the wastebasket. By the time Brian discovers what has happened, the janitor has cleaned up the entire dormitory. The contents of the wastebasket have been dumped into one of nine fully packed garbage cans outside the dorm. Brian insists that his roommate find the passport. Find the expected value of the number of garbage cans Brian's roommate will empty in order to find the passport.

Solution

When the roommate arranges the cans in a line for searching, the position of the can containing the passport is the only critical factor. Let X_i represent the outcome that the passport is in the ith can. It is reasonable to assume that each of the nine possible outcomes has probability 1/9. If the passport is in the ith can, then the roommate must empty i cans. Let the random variable R be defined by $R(X_i)=i$. The problem is solved by computing $EV(R)$. Now

$$\begin{aligned} EV(R) &= \sum_{i=1}^{9} i \Pr(R=a_i) \\ &= \sum_{i=1}^{9} i \Pr(X_i) \\ &= \sum_{i=1}^{9} i\left(\frac{1}{9}\right)=\left(\frac{1}{9}\right)\sum_{i=1}^{9} i=\left(\frac{1}{9}\right)45=5 \end{aligned}$$

so, on average, Brian's roommate can expect to search five cans.

Example 14 shows that there may be many applications of expected value when the random variable is measuring quantities other than money. Example 15 provides another example.

Example 15

In a study designed to test the efficiency of the postal service, a researcher mailed 1,000 letters from Los Angeles to New York on August 1. He kept a careful record of the delivery dates of each letter. The data are summarized in Table 10.2. What is the expected number of days for delivery of a letter?

Table 10.2

Date of delivery	Number of letters delivered
August 4	120
August 5	200
August 6	360
August 7	210
August 8	110

Solution

Formulate the question as a probability problem by letting the experiment consist of mailing a letter and observing the date of its delivery. Define the random variable R to be the number of days it takes the letter to be delivered. The problem is to find $EV(R)$.

From the data in Table 10.2, we see that R takes on values 3, 4, 5, 6, and 7 with respective probabilities of .12, .20, .36, .21, and .11. The expected value of R is $3(.12) + 4(.20) + 5(.36) + 6(.21) + 7(.11) = 4.99$. The researcher concluded that on average, the postal service takes just under 5 days to deliver a letter.

A final example shows how we may use expected value considerations in decision making.

Example 16

A suburban San Francisco construction firm is considering bidding on a contract to build one of two new schools. One possibility is that the firm will submit a bid to construct a high school. The firm estimates that it would make a $500,000 profit on the building, but that it would cost $10,000 to prepare the plans that must be submitted with the bid. (In estimating the profit, all costs, including that of the bid, have been considered.) The second possibility is a bid on a new elementary school. The firm has built several elementary schools in the recent past and estimates that the preparation costs for a bid would be only $5,000 while the potential profit is $400,000. The construction company has enough resources to submit only one bid. Past experience leads the company to estimate that it has one chance in five of submitting the winning bid for the high school and one chance in four for the winning bid on the elementary school. Which bid should the company prepare and submit?

Solution

The relevant data are summarized in Table 10.3.

Table 10.3

Contract	Profit	Bid cost	Probability of winning
High School	$500,000	$10,000	.20
Elementary School	$400,000	$5,000	.25

If the company submits a winning bid on the high school, its profit is $500,000. If it submits a bid on the high school that is not accepted, then its profit is $-\$10{,}000$. Thus, the expected value of submitting a bid on the high school is $(\$500{,}000)(.20) + (-\$10{,}000)(.80) = \$92{,}000$. The expected value for submitting a bid on the elementary school is $(\$400{,}000)(.25) + (-\$5{,}000)(.75) = \$96{,}250$. This indicates that the firm should submit the bid for constructing the elementary school, as it has a higher expected profit.

F. Variance and Standard Deviation

The expected value of a random variable R provides information about the "long-term" average value of R when the associated experiment is repeated over and over again. For many purposes, however, an average value may give insufficient or even misleading information.

Consider the distribution of income among a large population. A study shows that the average annual income per person in the United States is $30,000. Based on this figure, the Congress decides to classify all communities into three categories of income: below average, average, and above average. Communities in which the average income is below $30,000 will be singled out for financial assistance. Three hypothetical communities are of interest here, each having a population of 100 people. In the town of New Haven, every person earns exactly $30,000. In Bristol, 99 persons are unemployed and earn nothing, and one person has a trust fund that provides him with $3 million each year. In Ferrisburg, the income distribution is described by Table 10.4.

In each of the three communities, the total community income is $3 million, so the average income in each place is $30,000. The town of Bristol would be ineligible for the governmental assistance, even though 99 percent of the population is in a desperate situation! If we want to determine which communities need assistance, more information than average income is required. A measure of "deviation" from the average provides such additional data.

Table 10.4 Distribution of income in Ferrisburg

Income	Number of persons
$15,000	10
$24,000	20
$30,000	40
$36,000	20
$45,000	10

Table 10.5

i	$\Pr(R = i)$	$\Pr(S = i)$
1	.3	0
2	.05	.05
3	.05	.2
4	.2	.5
5	.05	.2
6	.05	.05
7	.3	0

To develop a measure of deviation, consider another example. There are two random variables, R and S, defined on the same sample space and each takes on the values 1, 2, 3, 4, 5, 6, 7, but with different probabilities. These probabilities are given in Table 10.5.

It is easy to calculate that $EV(R) = EV(S) = 4$. Both random variables have an average of 4. Yet it is more likely that the random variable S will take on values closer to the average than that R will. For example, the probability that R lies within 1 unit of the average value of 4 is

$$\Pr(R = 3, 4, 5) = .05 + .2 + .05 = .3,$$

while the probability that S lies within one unit of the average is

$$\Pr(S = 3, 4, 5) = .2 + .5 + .2 = .9.$$

In only about 1 time in 10 will the values of S differ from the mean by more than one unit, but this will happen about 7 out of 10 times for R. The random variable R has more "variability" or "deviation" about its average value than does the random variable S.

Suppose that an experiment is carried out using this sample space and the outcome results in the random variable R taking on the value i. Then the number

$$i - EV(R)$$

is called the *deviation of i from* $EV(R)$. The deviation will be positive if $i > EV(R)$ and negative if $i < EV(R)$.

In our example, $EV(R) = 4$, so that the deviations look like

i	-4
1	-3
2	-2
3	-1
4	0
5	1
6	2
7	3

The *sum* of all the deviations is not a good measure of the variation of the random variable, because that sum is 0. We would hope that the variation should be 0 only if the random variable always assumed its expected value, and that the variation would be positive otherwise. We could make the deviations positive in a variety of ways: consider only the absolute values $|i - EV(R)|$ or the square of the differences$(i - EV(R))^2$, for example. It turns out to be more convenient to use the squares of the deviations. [Recall the discussion of the least squares approach in Chapter 5.]

In constructing a measure of variation, then, we might simply add up the squares of the deviations from the expected values. This is not quite satisfactory either, since the sum would be the same, for example, for both random variables R and S in the case. The measure of variation should indicate that R varies more than S from the average values of 4. To obtain such a measure, multiply each particular $(i - EV(R))^2$ by the relative frequency with which it is likely to occur, $\Pr(R = i)$.

If this is done for the random variable R, the result is

$$\begin{aligned}&(1-4)^2\Pr(R=1) + (2-4)^2\Pr(R=2) + \cdots + (7-4)^2\Pr(R=7)\\ &\quad = 9(.3) + 4(.05) + 1.(.05) + 0(.2) + 1(.05) + 4(.05) + 9(.3) = 5.09.\end{aligned}$$

A similar computation for the random variable S yields

$$\begin{aligned}&(1.-4)^2\Pr(S=1) + \cdots + (7-4)^2\Pr(S=7)\\ &\quad = 9(0) + 4(.05) + 1(.2) + 0(.5) + 1(.2) + 4(.05) + 9(0) = .8\end{aligned}$$

so that the random variable S has a smaller measure of variability than the random variable R. Such a measure can be defined for any random variable.

DEFINITION Let Pr be a probability measure on a finite set E, and suppose R is a random variable taking on values $a_1, a_2, \ldots, a_k$. Then the *variance of* R, denoted $\mathrm{Var}(R)$ is the number

$$\begin{aligned}Var(R) &= (a_1 - EV(R))^2 Pr(R = a_1) + (a_2 - EV(R))^2 Pr(R = a_2) + \cdots\\ &\quad + (a_k - EV(R))^2 Pr(R = a_k) = \sum_{i=1}^{k} (a_1 - EV(R))^2 Pr(R = a_i)\end{aligned}$$

Example 17

Let R be the random variable whose values are the number of dots showing on the top of a fair die when it comes to rest after being rolled. As noted earlier, R takes on the values 1, 2, 3, 4, 5, 6, each with probability 1/6. The expected value of R is 7/2 (Example 11). The variation of R is

$$
\begin{aligned}
Var(R) &= \sum_{i=1}^{6}\left(i-\frac{7}{2}\right)^2\left(\frac{1}{6}\right) = \frac{1}{6}\sum_{i=1}^{6}\left(i=\frac{7}{2}\right)^2 \\
&= \frac{1}{6}\left[\left(1-\frac{7}{2}\right)^2+\left(2-\frac{7}{2}\right)^2+\left(3-\frac{7}{2}\right)^2+\left(4-\frac{7}{2}\right)^2+\left(5-\frac{7}{2}\right)^2+\left(6-\frac{7}{2}\right)^2\right] \\
&= \frac{1}{6}\left[\left(-\frac{5}{2}\right)^2+\left(-\frac{3}{2}\right)^2+\left(-\frac{1}{2}\right)^2+\left(\frac{1}{2}\right)^2+\left(\frac{3}{2}\right)^2+\left(\frac{5}{2}\right)^2\right] \\
&= \frac{1}{6}\left[\frac{25+9+1+1+9+25}{4}\right] = \frac{35}{12}
\end{aligned}
$$

Given a random variable R associated with a sample space, we may define a new random variable D_R, which takes on value $(i - EV(R))^2$ whenever R takes on value i. Then the variation of R is just the expected value of D_R. Another formula for the variation of R is given by the following formula.

THEOREM 4

$$Var(R) = \left[\sum_{i=1}^{k} a_i^2 Pr(R=a_i)\right] - [EV(R)]^2$$

Proof of Theorem 4 Expand the indicated sum in the definition of variation:

$$
\begin{aligned}
Var(R) &= \sum_{i=1}^{k}(a_i - EV(R))^2 Pr(R=a_i) \\
&= \sum_{i=1}^{k}\left[a_i^2 - 2a_i EV(R) + (EV(R))^2\right]Pr(R=a_i) \\
&= \sum_{i=1}^{k} a_i^2 Pr(R=a_i) - 2EV(R)\sum_{i=1}^{k} a_i Pr(R=a_i) + (EV(R))^2\sum_{i=1}^{k} Pr(R=a_i) \\
&= \sum_{i=1}^{k} a_i^2 Pr(R=a_i) - 2EV(R)EV(R) + [EV(R)]^2(1) \\
&= \sum_{i=1}^{k} a_i^2 Pr(R=a_i) - [EV(R)]^2
\end{aligned}
$$

The formula of Theorem 4 is easier to use than the definition of variation since the former requires only one subtraction while the latter demands k subtractions. Using the

formula to compute the variation of the random variable to Example 17, for instance, involves the calculation

$$Var(R) = \frac{1}{6}\left[1^2 + 2^2 + \cdots + 6^2\right] - \left(\frac{7}{2}\right)^2 = \frac{91}{6} - \frac{49}{4} = \frac{35}{12}$$

Since the variance is calculated using squares of the values of the random variable, the units of $Var(R)$ are the squares of the units of R. If the values of R are dollars, then the units of $Var(R)$ would be "square dollars." In many instances, it is convenient to have a measure of variability about the expected value that is in the same type of units as the random variable itself. This can be accomplished by determining the nonnegative square root of the variance. The resulting number, $\sqrt{Var(R)}$, is called the *standard deviation* of R. It is often denoted by $SD(R)$.

As a final example, note that the standard deviation of the random variable associated with the throwing of a fair die is $\sqrt{35/12}$, which is approximately 1.71.

III. A Probabilistic Model

Chapter 3 discussed a deterministic model for single species population growth, the so-called pure birth process. The assumptions of this process are that the population is made up entirely of identical organisms reproducing independently at a rate that is the same for every individual at all moments. The deterministic model for the pure birth process is the first-order differential equation $dN/dt = bN$, where $N = N(t)$ is the population at time t and b is the positive constant birth rate for each individual. The solution of the differential equation is $N(t) = Ae^{bt}$ where A is the population at time $t = 0$.

The deterministic model assumes not simply that each individual *may* reproduce but that in actuality it *does* reproduce with absolute certainty. This section outlines a *probabilistic model* for the pure birth process. The assumption that makes this a probabilistic model is the assertion that there is a certain probability that a particular individual will reproduce in a given time interval.

More precisely, we assume that the probability of reproduction in a very short time interval is directly proportional to the length of the interval—that is, there is a constant b such that in any small time interval of duration Δt the probability of reproduction is $b\Delta t$. Take Δt as so small that no individual can reproduce more than once in the time interval. Thus, during the interval of length Δt, a given individual either produces one offspring with probability $b\Delta t$ or produces no offspring with probability $1 - b\Delta t$. In a population of N organisms, the probability of a birth during the time interval is $Nb\Delta t$.

Let $P_N(t + \Delta t)$ denote the probability that the population is of size N at time $t + \Delta t$. This outcome can occur in one of two distinct ways:

(a) At time t, there were $N - 1$ individuals and one birth occurred in the next Δt seconds.

(b) At time t, there were N individuals in the population and no births occurred in the next Δt seconds.

(By choosing a small enough Δt, it is safe to assume that not more than one birth takes place. As in any pure birth process, the assumption is that no individual dies.)

For each positive integer $N(N = 1, 2, 3, \ldots)$, the fact that (a) and (b) describe disjoint events gives

$$P_N(t + \Delta t) = P_{N-1}(t)b(N - 1)\Delta t + P_N(t)(1 - b)N\Delta t \tag{1}$$

Rewrite this equation as

$$P_N(t + \Delta t) - P_N(t) = -bN\Delta t P_N(t) + P_{N-1}(t)b(N - 1)\Delta t$$

and divide each side by Δt to obtain

$$\frac{P_N(t + \Delta t) - P_N(t)}{\Delta t} = -bNP_N(t) + P_{N-1}(t)b(N - 1) \tag{2}$$

Taking the limit of each side of (2) as Δt tends to zero yields a differential equation:

$$\frac{dP_N(t)}{dt} = -bNP_N(t) + b(N - 1)P_{N-1}(t) \tag{3}$$

There is such a differential equation for each positive value of N. Denote the size of the population at time 0 by A so that $P_A(0) = 1$ and $P_N(0) = 0$ whenever $N \neq A$.

When N is equal to the original population A, Eq. (3) becomes

$$\frac{dP_A(t)}{dt} = -bAP_A(t) + b(A - 1)P_{A-1}(t) \tag{4}$$

Under the assumption of this simple model that there are only births and no deaths, the population is always at least as large as A. The probability that there are ever fewer than A individuals is 0. In particular, $P_{A-1}(t) = 0$ for all t. Eq. (4) then simplifies to

$$\frac{dP_A(t)}{dt} = -bAP_A(t) \tag{5}$$

If we let $y = P_A(t)$, Eq. (5) is of the form $dy/dt = -bAy$, which can be solved by integration to obtain $y = y_0 e^{-bAt}$, where $y_0 = y(0) = P_A(0) = 1$. Thus, the model gives

$$P_A(t) = e^{-bAt}. \tag{6}$$

Eq. (6) predicts the probability that the population is still at size A at time t—that is, the probability that no births have occurred in the interval $[0,t]$. Note that this probability is always positive, but that it decreases as time increases, asymptotically approaching 0 as t increases without bound.

Thus far, the consequences derived from this model are

$$P_N(t) = \begin{cases} 0 & \text{if } N < A \\ e^{-bAt} & \text{if } N = A \end{cases}$$

The next step is to calculate $P_{A+1}(t)$, which is the probability of a population of $A+1$ individuals at time t. Substitute $N=A+1$ into Eq. (3):

$$\frac{dP_{A+1}(t)}{dt} = -b(A+1)P_{A+1}(t) + bAP_A(t)$$

and use the result of Eq. (6) to obtain

$$\frac{dP_{A+1}(t)}{dt} + b(A+1)P_{A+1}(t) = bAe^{-bAt} \tag{7}$$

Eq. (7) has the form

$$\frac{dx}{dt} + b(A+1)x = bAe^{-bAt} \tag{8}$$

where $x=P_{A+1}(t)$. Eq. (8) is a first-order linear differential equation. It may be solved (see Appendix V) by multiplying through by an integrating factor, $e^{b(A+1)t}$, then integrating:

$$e^{b(A+1)t}\frac{dx}{dt} + e^{b(A+1)t}b(A+1)x = bAe^{-bAt}e^{b(A+1)t} = bAe^{bt} \tag{9}$$

or

$$\frac{d}{dt}e^{b(A+1)t}x = bAe^{bt} \tag{10}$$

which becomes, upon integration,

$$e^{b(A+1)t}x = Ae^{bt} + \text{Constant} \tag{11}$$

which is to say

$$e^{b(A+1)t}P_{A+1}(t) = Ae^{bt} + \text{Constant}. \tag{12}$$

Since $P_{A+1}(0)=0$, the constant of integration is equal to $-A$ and the solution of the differential equation is

$$P_{A+1}(t) = Ae^{-Abt}(1-e^{-bt}) \tag{13}$$

Once $P_{A+1}(t)$ is known, it can be used to find $P_{A+2}(t)$ by making use of Eq. (13) and the modeling Eq. (3) with $N=A+2$:

$$\frac{dP_{A+2}(t)}{dt} = -b(A+2)P_{A+2}(t) + b(A+1)P_{A+1}(t) \tag{14}$$

or

$$\frac{dP_{A+2}(t)}{dt} + b(A+2)P_{A+2}(t) = b(A+1)Ae^{-Abt}(1-e^{-bt}) \tag{15}$$

Eq. (15) is again a first-order linear differential equation. Multiply each side of the equation by $e^{b(A+2)t}$, integrate, and then use the fact that $P_{A+2}(0) = 0$ to obtain the solution

$$P_{A+2}(t) = \frac{(A+1)A}{2} e^{-Abt}(1-e^{-bt})^2 \tag{16}$$

Continue in a similar manner to find $P_{A+3}(t)$, $P_{A+4}(t)$, $P_{A+5}(t)$, and so forth. The general formula, which may be checked by induction, is

$$P_N(t) = \binom{N-1}{A-1} e^{-Abt}(1-e^{-bt})^{N-A} \text{for all } N \geq A \tag{17}$$

where

$$\binom{N-1}{A-1} = \frac{(N-1)!}{(A-1)!(N-A)!}$$

(see Exercise 58).

The solution (Eq. (17)) of the probabilistic model for a pure birth process gives the probability distribution of the size of the population at time t. While the deterministic model gives a single number as the prediction for the population size at time t, the probabilistic model gives much more information—namely, the relative likelihood of each different possible population size at time t.

The deterministic model was much simpler to treat mathematically than the probabilistic one. What is the connection between these two models? In what sense is the deterministic model an approximation for the probabilistic one? This relationship becomes clearer if the solution of the probabilistic model is used to compute the *expected value* of the size of the population at time t. This expected value turns out to be Ae^{bt}, the number predicted by the deterministic model. The probabilistic model also provides a measure of the variation from this expected value, a measure that is unavailable if a deterministic approach alone is used. The variation is $Ae^{bt}(e^{bt}-1)$. The calculation of expected value and variance is left to the exercises.

IV. Stochastic Processes

A. Definitions

A *stochastic process* is a sequence of experiments in which the outcome of each experiment depends on chance. A stochastic process consisting of a finite number of experiments, each having a finite number of possible outcomes is called a *finite stochastic process*. (The Greek word "stochos" means "guess.")

The experiments may or may not be related to each other. The outcomes of one experiment may or may not affect the probabilities of the outcomes of subsequent

experiments. We will emphasize two types of stochastic processes in this book. In the first, the experiments are mutually independent. In the second, called a *Markov chain*, the likelihood of an outcome of one experiment depends only on the outcome of the immediately preceding experiment.

Stochastic processes have been widely used as mathematical models in the study of many diverse social, physical, and biological problems. Researchers have used stochastic processes to investigate problems in economics, genetics, learning theory, educational planning, demography, job mobility, social conformity, evolution, consumer buying behavior, the geographical diffusion of innovations, and in many other fields. In many of these applications a process is studied that may be in one of various "states" at each moment. The "states" correspond to the outcomes of the experiments. A sequence of experiments is constructed by examining the process at equally spaced time intervals.

Example 18

The Board of Trustees of a small Vermont college decides to choose a student from one of two dormitories to serve on a housing committee. A dorm will be chosen at random and then a student will be selected at random from that dorm. The dormitories are Starr Hall and Forest Hall. There are 30 students in Starr; 20 oppose coeducational housing and 10 favor it. Of the 60 students living in Forest, only 10 favor coed housing, and all the others oppose it. What is the probability that the student chosen to serve on the committee will favor coed housing?

Solution

Describe the situation in terms of a stochastic process involving two experiments, each with two possible outcomes. In experiment 1, a dorm is chosen. There are two possible outcomes: Starr (S) and Forest (F), with probabilities $\Pr(S) = \Pr(F) = 1/2$. In experiment 2, a student is chosen. The possible outcomes are: Approves coed housing (A) or disapproves coed housing (D). The probabilities of A and D depend on which dorm is chosen—that is, they are conditional probabilities. These probabilities are

$$\Pr(A|S) = \frac{10}{30} = \frac{1}{3}, \quad \Pr(D|S) = \frac{20}{30} = \frac{2}{3}$$

$$\Pr(A|F) = \frac{10}{60} = \frac{1}{6}, \quad \Pr(A|F) = \frac{50}{60} = \frac{5}{6}$$

The problem is to determine $\Pr(A)$. By Theorem 1 and the definition of conditional probability,

$$\begin{aligned}
\Pr(A) &= \Pr(A \cap F) + \Pr(A \cap F^C) \\
&= \Pr(A \cap F) + \Pr(A \cap S) \\
&= \Pr(A|F)\Pr(F) + \Pr(A|S)\Pr(S) \\
&= \left(\frac{1}{6}\right)\left(\frac{1}{2}\right) + \left(\frac{1}{3}\right)\left(\frac{1}{2}\right) = \frac{1}{4}
\end{aligned}$$

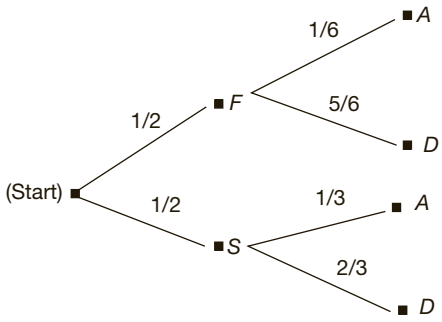


Fig. 10.1 Tree diagram corresponding to Example 18.

B. Tree Diagrams

A convenient way to study stochastic processes is through the use of tree diagrams. A tree diagram summarizing the information of Example 18 is shown in Fig. 10.1.

From the starting point, or *node*, there are two *branches*, corresponding to the two possible outcomes of the first experiment. The numbers along each branch give the probabilities of each outcome. From each outcome of the first experiment there are again a pair of branches representing the two possible outcomes of the second experiment. The numbers on these branches indicate the probabilities of the outcomes.

The probability of tracing through any particular path in a tree diagram is the product of the probabilities along the branches of that path. To find the probability of selecting a student who approves coed housing, $\Pr(A)$, simply add up the probabilities of all distinct paths from the start to outcome A. In a similar fashion, the probability of a particular outcome of the final experiment of a sequence can be computed by summing the probabilities of every path in the tree diagram, which ends at that outcome. The next example gives an additional illustration.

Example 19

The winners of some tennis matches are determined by playing a best-of-three sets competition. The competitors keep playing until one of them wins two sets; no set may end in a tie. Fig. 10.2 shows a tree diagram illustrating the possible outcomes for one of the players. Note that the outcomes of certain sets determine whether or not successive sets are played.

Suppose the player under study has an even chance of winning the first set, that whenever she wins a set, she has a tendency to relax in the next set so that her probability of winning drops to 3/8, and that whenever she loses a set, she exerts herself to such an extent that her probability of winning the next set jumps to 3/4. What is the probability that she will win the match?

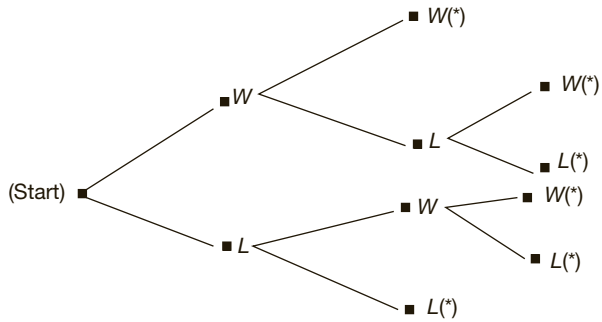


FIGURE 10.2 Tree diagram for best-of-three tennis competition. W = Win, L = Lose, and (*) indicates that the match terminates at the node.

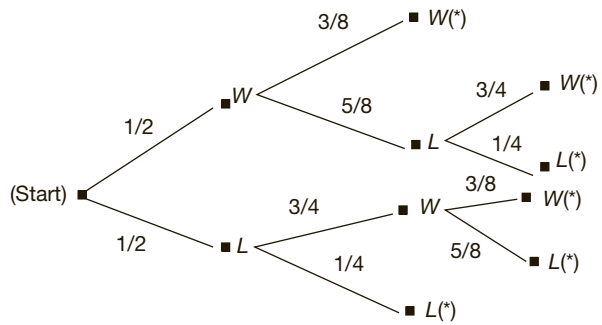


FIGURE 10.3

The probabilities along the branches given by this information are indicated in Fig. 10.3.

The probability of winning is

$$\begin{aligned}\Pr(\text{Wins match}) &= \Pr(WW) + \Pr(WLW) + \Pr(LWW) \\ &= \left(\frac{1}{2}\right)\left(\frac{3}{8}\right) + \left(\frac{1}{2}\right)\left(\frac{5}{8}\right)\left(\frac{3}{4}\right) + \left(\frac{1}{2}\right)\left(\frac{3}{4}\right)\left(\frac{3}{8}\right) = \frac{9}{16}\end{aligned}$$

where $\Pr(LWW)$ is the probability of losing the first set and then winning the second and third sets. The events WW and WLW are similarly defined. After the first set, the probability of winning a subsequent set depends only on the result of the previous set. This probability does not depend, for example, on the total number of sets she has won previously, or on how many sets have been played. The assumptions about this player are those of a Markov chain.

As an alternative possibility, suppose the results of each set are independent of the results of earlier sets. Then the probabilities along each branch might be assigned values of 1/2. In this case, the probability of winning the match is given by

$$\Pr(WW) + \Pr(WLW) + \Pr(LWW) = \frac{1}{2}\frac{1}{2} + \frac{1}{2}\frac{1}{2}\frac{1}{2} + \frac{1}{2}\frac{1}{2}\frac{1}{2} = \frac{1}{2}$$

As a third possibility, suppose the player gets stronger as the match goes on so that her probability of winning each set is twice the probability of winning the previous set. If her probability of winning the first set is $\frac{1}{5}$, then her probability of winning the match is

$$\Pr(WW) + \Pr(WLW) + \Pr(LWW) = \frac{1}{5}\frac{2}{5} + \frac{1}{5}\frac{3}{5}\frac{4}{5} + \frac{4}{5}\frac{2}{5}\frac{4}{5} = \frac{54}{125}$$

The assumptions in this case do not correspond to either a mutually independent sequence of experiments or to a Markov chain.

C. Examples

This chapter concludes with brief descriptions of a number of stochastic processes in which a single experiment is repeated a number of times. Each repetition of the experiment will be a "step," and the outcomes of the experiment are "states." The viewpoint is one of studying a process that moves in a sequence of steps through a set of states. At each step, there are certain probabilities of being in each of the possible states; these probabilities are the likelihoods of the various outcomes of the experiment.

Example 20

At each a step a coin is tossed. There are two possible states: heads and tails. The probability of being in the "head" state on each step is $1/2$. It is independent of all previous steps.

Example 21

Each step is a new day. The states are the hours of the day. We are interested in the time you go to sleep. There is a probability that can be attached to each of the 24 states for today's step. Knowledge of what the probabilities were yesterday will help determine what probabilities to assign for the states today.

Example 22

Each step is a new month. There are two states, "Flakes No More" and "Head, Neck, and Shoulders," two antidandruff shampoos. We are concerned with the percentage of consumers who use each product. Interpret this as the probability associated with picking a person at random and determining which shampoo she uses. If 60% of the consumers use Flakes No More, assume that there is a probability of .6 that a randomly chosen person uses that shampoo.

Example 23

An investigator for the Equal Opportunity Commission analyzed job mobility for women in Cook County, Illinois. She determined from census data the percentage of women who were professionals, skilled workers, and unskilled workers. She amassed data for six successive generations of women. She then formulated a stochastic process with six steps and three states. Each step corresponded to a new generation. The states were: professional, skilled, and unskilled.

Example 24

A particle is constrained to move along the x-axis. It starts at 0 and at each step it may move one unit to the right or one unit to the left. The direction of motion is randomly chosen. To view this motion as a stochastic process, consider that the particle may be in any one of an infinite number of states, corresponding to possible positions it might reach on the x-axis $(0, \pm 1, \pm 2, \ldots)$. The study of random walk along a line, in the plane, or in higher dimensional spaces has many applications in modern physics.

Example 25

A mathematically inclined sports broadcaster applied stochastic processes to study the movement of the puck in a hockey game between the Montréal Canadians and the Philadelphia Flyers. The playing area of the hockey rink can be divided into five states: center ice, Montréal territory, Philadelphia territory, Montréal goal, and Philadelphia goal. Each step corresponds to a change of state of the puck. Thus, the puck cannot enter the state of "Montréal goal" from the state of "center ice" without first passing through the state of "Montréal territory."

Example 26

The simplest type of inheritance of traits in human beings occurs when a trait is governed by a pair of genes, each of which may be of two types, say *A* and *B*. An individual may have an *AA* combination or *AB* (which is genetically the same as *BA*) or *BB*. Very often, the *AA* and *AB* types are indistinguishable in appearance, in which case it is said that *A dominates B*.

An individual is called *dominant* if he has *AA* genes, *recessive* if he has *BB*, and *hybrid* if he has an *AB* pairing.

In reproduction, the offspring inherits one gene of the pair from each parent. A basic assumption of genetics is that these genes are selected at random, independently of each other.

Geneticists are interested in the probability of an individual being in one of the three states—dominant, recessive, or hybrid—and in how this probability changes from generation to generation. Each succeeding generation is a new step in this stochastic process.

The next chapter is devoted to a detailed study of Markov chains, the type of stochastic process that has been most widely used in model building in the social and life sciences.

EXERCISES

I. The Need For Probability Models

1. Can you think of any further objections to the use of deterministic models in the social or life sciences?

2. Are the responses to the objections made against deterministic models adequate to your way of thinking?

3. In what ways can physics be validly described as more "rigorous" than mathematical social science?

II. What Is Probability?

A. Fundamental Definitions

4. Find a probability measure consistent with the observation that a flipped coin shows tails eight times more frequently than heads.

5. **(a)** Show that the definition of $\Pr(X)$ in Example 3 establishes a valid probability measure.

(b) Show that the definition of $\Pr(X)$ in Example 4 establishes a valid probability measure.

6. Construct an example of a sample space on a set E so that for some nonempty subset X of E, $\Pr(X) = 0$.

7. Prove Theorem 1.

8. Find $\Pr(A \cup B)$ if $\Pr(A) = .6$, $\Pr(B) = .7$, and $\Pr(A \cap B) = .5$.

9. If $\Pr(A \cap B) = 1/5$, $\Pr(A^C) = 1/4$, and $\Pr(B) = 1/3$, find $\Pr(A \cup B)$.

10. Prove that $\Pr(X \cup Y \cup Z) = \Pr(X) + \Pr(Y) + \Pr(Z) - \Pr(X \cap Y) - \Pr(X \cap Z) - \Pr(Y \cap Z) + \Pr(X \cap Y \cap Z)$ for any three subsets X, Y, Z of a sample space.

11. Roll a pair of dice and assume that all 36 possible outcomes are equiprobable. Find the probability of each of the following events:

(a) The sum of the numbers on the faces is 7.

(b) The sum of the numbers is odd.

(c) Both numbers are odd.

(d) The product of the numbers is greater than 10.

12. The *odds* in favor of an outcome are r to s if the probability of the outcome is p and $r/s = p/(1-p)$.

(a) Show that if the odds in favor of an outcome are r to s, then the probability that the outcome will occur is $r/(r+s)$.

(b) If you believe the Los Angeles Dodgers have 1 chance in 3 of winning the World Series next year, what odds should you offer in making bets about the series?

13. Probability measures may also be defined on some infinite sets. Let E be the set of all positive integers and let the weight of integer j be equal to $(1/2)^j$. The probability, $\Pr(X)$, of a subset X of E is defined to be the sum of the weights of the elements of X.

(a) Show that $\Pr(E) = 1$.

(b) Show that Pr, defined in this manner, satisfies the defining properties of a probability measure.

(c) What is the probability that an integer chosen from this sample space will be even?

B. Conditional Probability

14. Under what conditions does $\Pr(X|Y) = \Pr(Y|X)$?

15. In Example 5, find the probability that Mazzoli wins given that Levine loses. Which candidate benefits most from Levine's withdrawal? (Note that most may be defined in several different ways.)

16. There are three chests, each having two drawers. The first chest has a gold coin in each drawer, the second chest has a gold coin in one drawer and a silver coin in the other, and the third chest has a silver coin in each drawer. A chest is chosen at random and a drawer opened. If that drawer contains a gold coin, what is the probability that the other drawer contains a gold coin? Warning: The answer is not 1/2.

17. If X and Y are events with positive probabilities, show that

$$\Pr(X|Y)\Pr(Y) = \Pr(Y|X)\Pr(X)$$

18. Consider the following problem and three proposed solutions. Problem: A census taker is interviewing Mr. Linovitz who is babysitting his son. Mr. Linovitz tells the census taker that he has two children. What is the probability that the other child is a boy?

Solution 1 There are four possibilities. Mr. Linovitz has two sons, he has two daughters, he had a son first and then a daughter, or he had a daughter and then a son. In only one of the four cases is the other child also a boy. Thus, the probability is 1/4.

Solution 2 Since we know that one of the children is a boy, there are only three possibilities: two sons, son first and then daughter, daughter first and then son. The probability is 1/3.

Solution 3 There are only two possibilities: the other child is either a boy or a girl. The probability is 1/2.

Which of the three solutions is correct?

C. Bayes's Theorem

19. There are three cookie jars in our kitchen containing both chocolate and vanilla cookies. The green jar contains 3 chocolate and 4 vanilla cookies, the blue jar contains 5 chocolate and 2 vanilla, and the red jar contains 2 chocolate and 5 vanilla cookies. While his parents are asleep, Eli sneaks downstairs to the darkened kitchen and steals a cookie. After biting it, he discovers the cookie is a chocolate one. What is the probability that it came from the blue jar?

20. A recently devised self-administered test for pregnancy has been found to have the following reliability. The test detects 75% of those who are actually pregnant, but does not detect pregnancy in 25% of this group. Among those women who are not pregnant, the test detects 85% as not being pregnant, but indicates 15% of this group as being pregnant. It is known that in a large sample of college women 2% are pregnant. Suppose a coed is chosen at random, given the test, and registers as being pregnant. What is the probability that she actually is?

21. If Detective Friday (Example 7) never takes drugs, but is tested every week, what is the probability that he will have at least one positive test in the first 6 months?

22. The National Cancer Institute estimates that 1 in 5,000 women in the United States has invasive cervical cancer. One of the major diagnostic tests used today is the Pap smear, which has a specificity of .95 and a sensitivity somewhere between 70% and 80%. What is the probability that a randomly chosen woman who gets a positive test actually has cancer? If a cancer-free woman has an annual Pap smear beginning at age 20, how likely is she to receive at least one false-positive test by the time she reaches middle age?

D. Independent Events

23. A dog has a litter of four puppies. Is it more likely that there will be two males and two females or exactly three of the same gender?

24. Sandra and Matt independently proofread the manuscript of Ron's book. Sandra found 60 errors and Matt identified 40, 20 of which Sandra had also reported.

(a) What is the total number of distinct errors they discovered?

(b) Estimate the number of errors that remain unnoticed.

25. Suppose you're on the game show *Let's Make a Deal*, and you're given the choice of three doors. Behind one door is a car, the others, goats. You pick a door, say #1, and the host, Monty Hall, who knows what's behind the doors, opens another door, say #3, which has a goat. He says to you: "Do you want to pick door #2?" Is it to your advantage to switch your choice of doors?

26. John, who is a better player than you, offers you the opportunity to play a sequence of three games where you will play alternately between John and Pete, who is an even better player than John. You will receive a large cash prize if you win at least two games in a row. You have a choice between playing Pete in games 1 and 3, facing John in game 2 or playing John in the first and last games with Pete as your opponent in the second game. Suppose your probability of beating John in a single game is q and of defeating Pete is p. Who would you chose to face in game 1?

27. If X and Y are disjoint subsets of E, under what conditions are they independent events?

28. An urn contains eight marbles numbered from 1 to 8. A marble is picked at random and removed from the urn. Then a second marble is picked at random from the remaining seven marbles.

(a) Find the probability that the numbers on the two chosen marbles differ by two or more.

(b) What is the answer to (a) if the first marble is replaced in the urn before the second marble is chosen?

29. *Polya's urn scheme*. An urn originally has r red marbles and b black marbles. A marble is selected at random and removed from the urn. Then that marble and c other marbles of the same color are added to the urn. This procedure is repeated $n - 1$ additional times. Show that the probability of selecting a red ball at any trial is $r/(b + r)$.

30. Suppose X, Y, and Z are mutually independent events and $\Pr(X \cap Y) \neq 0$. Show that $\Pr(Z|X \cap Y) = \Pr(Z)$.

31. The *factorial* of a positive integer n is denoted $n!$ and is defined to be the product of the integers from 1 to n. Thus, $3! = 3 \times 2 \times 1 = 6$. For convenience, we define $0! = 1$.

(a) Compute 4!, 5!, 6!.

(b) Show that $(n+1)! = (n+1)n!$ for any positive integer n.

32. The symbol $\binom{n}{k}$ where n and k are nonnegative integers and $k \leq n$ is defined to be the number $\frac{n!}{k!(n-k)!}$

(a) Compute $\binom{6}{k}$ for $k = 0, 1, 2, 3, 4, 5, 6$.

(b) Show that $\binom{n}{k} = \binom{n}{n-k}$.

(c) Prove that $k\binom{n}{k} = n\binom{n-1}{k-1}$.

(d) Prove that $\binom{n}{k}$ is always an integer.

33. **(a)** Show that the number of distinct ways of arranging r objects in a straight line is $r!$.

(b) Show that the number of distinct ways of choosing k objects from a set of n objects is $\binom{n}{k}$.

(c) An experiment has two possible outcomes: a success with probability p and a failure with probability $1 - p$. Show that the probability of obtaining exactly k successes in n repetitions of the experiment is

$$\binom{n}{k} p^k (1-p)^{n-k}$$

34. **(a)** Show that $(p+q)^n = \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k}$ if n is a positive integer.

(b) Prove that $\sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} = 1$.

E. Expected Value

35. In Example 16, let p and q be the probabilities of winning the bids for the high school and the elementary school, respectively. For which values of p and q are the expected values of the two bids equal?

36. In the game of "craps" a player rolls a pair of dice. If the sum of the numbers shown is 7 or 11, he wins. If it is 2, 3, or 12, he loses. If it is any other sum, he must continue rolling the dice until he either repeats the same sum (in which case he wins) or he rolls a 7 (in which case he loses). Suppose the outcome of each round is a win or a loss of \$5. What is the probability that he will win a round? What is the expected value of shooting craps?

37. A roulette wheel has the numbers 0, 1, 2, ..., 36 marked on 37 equally spaced slots. The numbers from 1 to 36 are evenly divided between red and black. A player may bet on either color. If a player bets on red and a red number turns up after the wheel is spun, she receives twice her stake. If a black number turns up, she loses her stake. If 0 turns up, then the wheel is spun again until it stops on a red or a black. If this is red, the player receives only her original stake, and if it is black, she loses her stake. If a player bets \$1 on red with each spin, what is her expected value of winning?

38. A new-car dealer receives nine station wagons from the factory for sale. The next day, a telegram from the factory informs him that there is strong reason to believe that the brakes in two of the cars are defective. What is the expected value of the number of cars the dealer will have to test in order to find both defective ones?

39. **(a)** Show that $\sum_{k=0}^{n} k\binom{n}{k} p^k (1-p)^{n-k} = n \sum_{k=1}^{n} \binom{n-1}{k-1} p^k (1-p)^{n-k}$.

(b) Show that the sum in (a) is also equal to $np \sum_{k=1}^{n-1} \binom{n-1}{k} p^k (1-p)^{n-k-1}$.

(c) Show that the sum in (b) is equal to np.

(d) Find the expected number of heads in n tosses of a coin if the probability of a head on each toss is p.

40. Let R and S be random variables defined on a set E. Then the random variable $R + S$ is the function defined on E whose value is $(R+S)(e) = R(e) + S(e)$ for each element e of E. Prove that $EV(R+S) = EV(R) + EV(S)$.

41. Extend the result of Exercise 40 to show that if $R_1, R_2, \ldots, R_k$ are random variables on a set E, and $X = \sum_{i=1}^{k} Ri$ is defined by $X(e) = \sum_{i=1}^{k} Ri(e)$, then $EV(X) = \sum_{i=1}^{k} EV(Ri)$.

42. An experiment has two possible outcomes: success with probability p and failure with probability q. Suppose the experiment is repeated n times in such a

manner that the outcomes on successive repetitions are independent. Let Ri be the random variable with the value 1 if the outcome of the ith experiment is a success and 0 if it is a failure. Show that $EV(Ri) = p$ and $EV(X) = np$ if $X = \sum_{i=1}^{n} Ri$. Find that the expected number of heads in n tosses.

43. Let R be a random variable defined on a set E. If c is any constant, define the random variable cR and prove that $EV(cR) = cEV(R)$.

44. Let R be a random variable with nonnegative values $a_1, a_2, a_3 \ldots$, defined on the sample space of Exercise 13. The *expected value* of R is defined to be $EV(R) = \sum_{i=1}^{\infty} a_i \Pr(R = a_i)$ provided this infinite series converges.

(a) Suppose $R(j) = 3^{-j}$ for each j in E. Compute $EV(R)$.

(b) Suppose $R(j) = 2^j$ for each j in E. Does R have an expected value? Interpret this result in the light of the St. Petersburg paradox.

(c) How would you define expected value for a random variable defined on the set of all integers greater than or equal to a fixed positive integer A?

45. A coin has probability p of Heads. Show that the expected number of flips until you get a Heads is $1/p$.

46. China attempted to implement a "One Child" policy whereby each family was limited to one child. Because of the traditional importance of having a son in Chinese culture, many couples were aborting female fetuses or putting baby girls up for adoption so that they might try again for a son. Suppose China changed to a "One Boy" policy whereby families could have as many children as they wanted until they had a boy who would then be their last child. What would be the expected family size under the "One Boy" policy? What would be the expected ratio of boys to girls in China a generation or two in the future? (Note that the large population of China would mean that there would be many families with 4 or 5 or 6 girls.)

F. Variance and Standard Deviation

47. Find the variance and standard deviation of income in New Haven, Bristol, and Ferrisburg.

48. A random variable takes on the values $-2, -1, 0, 1, 2$, with probabilities .2, .3, .3, 1, .1, respectively. Find the expected value, variance, and standard deviation.

49. Show that the variance of a random variable is zero if and only if the random variable takes on exactly one value with probability 1.

50. Toss a coin eight times and let R denote the number of heads. Find the expected value, variance, and standard deviation of R if

(a) The coin is a fair one.

(b) The coin is weighted so that it comes up heads with probability 3/5.

51. Show that the variance of a random variable can be determined from the formula $\mathrm{Var}(R) = EV(R^2) - (EV(R))^2$.

52. If R is a random variable and c is a constant, show that $\mathrm{Var}(cR) = c^2 \mathrm{Var}(R)$.

53. Suppose R is a random variable defined on a set E and b is a constant. Define a new random variable, $R + b$, by $(R + b)(e) = R(e) + b$ for each e in E. Show that $\mathrm{Var}(R) = \mathrm{Var}(R + b)$.

54. Find the variance in the number of heads in n tosses of a coin if the probability of a head on each toss is p. (See Exercise 39; you should arrive at the number $np(1 - p)$.)

55. Suppose R and S are random variables defined on the same sample space. What is the relation between $\mathrm{Var}(R + S)$, $\mathrm{Var}(R)$, and $\mathrm{Var}(S)$?

III. A Probabilistic Model

56. Verify the details in the derivation of $P_{A+2}(t)$.

57. Use $P_{A+2}(t)$ and Eq. (3) to compute $P_{A+3}(t)$.

58. Prove, by induction on N, that $P_N(t) = \binom{N-1}{A-1} e^{-bAt}(1 - e^{-bt})^{N-A}$ for each $N \geq A$.

59. Show that $P_N(t)$ induces a probability measure on the set of all integers greater than or equal to A. You must show that $\sum_{N=A}^{\infty} P_N(t) = 1$.

60. Graph $P_N(t)$ as a function of t. Show that, for a fixed $N > A$, $P_N(t)$ first increases and then decreases toward 0. For what value of t is the probability greatest?

61. In this problem, you will compute the value of the population at time t for the stochastic pure birth process. For convenience, let P_N denote $P_N(t)$.

(a) With this notation, show that Eq. (3) becomes

$$\frac{dP_N}{dt} = -bNP_N + b(N-I)P_{N-1}.$$

(b) Use the fact that $P_{A-1}(t) = 0$ for all t to show that

$$\sum_{N=A}^{\infty} \left(-N^2P_N + N(N-1)\right)P_{N-1} = \sum_{N=A}^{\infty} P_N(N+1)N - N^2 = \sum_{N=A}^{\infty} NP_N$$

(c) Let $E = E(t)$ denote the expected value of $P_N = P_N(t)$. Show that $E(0) = A$.

(d) Show that $E = \sum_{N=A}^{\infty} NP_N$.

(e) Justify each step in the following calculation:

$$\frac{dE}{dt} = \sum_{N=A}^{\infty} N(-bNP_N + b(N-1)P_{N-1}) = b\sum_{N=A}^{\infty}(-N^2P_N + N(N-1)P_{N-1}) = b\sum_{N=A}^{\infty} NP_N = bE$$

(f) The expected value E then satisfies the differential equation $\frac{dE}{dt} = bE$ with initial condition $E(0) = A$. Show that the solution of this equation is $E(t) = Ae^{bt}$.

62. Use the approach of Exercise 61 to show that the variance of $P_N(t)$ is given by $Ae^{bt}(e^{bt} - 1)$.

IV. Stochastic Processes

63. In Example 17, suppose the probability of choosing a dorm is proportional to the number of residents in it.

(a) Show that $\Pr(F) = 2/3$.

(b) Determine $\Pr(A)$.

64. Draw a tree diagram representing the possible outcomes of a baseball World Series. The winner of the series is the first team to win four games.

65. If a baseball team has a probability of .55 of winning each World Series game in which it competes, find

(a) The probability that it sweeps the series in four games

(b) The probability that it wins the series after losing the first two games

(c) The probability that it wins the series

SUGGESTED PROJECTS

1. The stochastic pure birth process may be generalized to a birth-and-death process. In addition to the basic assumptions of the pure birth process, suppose that the probability that an individual will die in a short time interval is directly proportional to the length of the interval. Show that this assumption leads to the equation

$$P_N(t+\Delta t) = P_{N-1}(t)b(N-1)\Delta t + P_N(t)(1 - (b+d)N\Delta t) + P_{N+1}(t)d(N+1\Delta t$$

for some positive constants b and d.

Derive a set of differential equations analogous to those of Eq. (3) of the text. Solve, if possible, the equations for $P_N(t)$, where $N = A, A \pm 1, A \pm 2, \ldots$

Show that the expected value of the population at time t is $Ae^{(b-d)t}$. Compare the probabilistic model with the deterministic one. Find the probability that a population governed by a birth-and-death process will eventually become extinct.

2. As a different generalization of the pure birth process, suppose the proportionality factor b is not constant, but is a function of the population N. Show that this leads to a model of the form

$$\frac{dP_N(t)}{dt} = -b_NP_N(t) + b_{N-1}P_{N-1}(t).$$

Investigate such models.

3. Develop in as much detail as possible a probabilistic model for logistic population growth.

4. Investigate how misconceptions of probability can affect legal decisions. These misunderstandings are often called the "prosecutor's fallacy" or the "defense attorney's fallacy" or even the "juror's fallacy." Some of the common errors involve confusing $\Pr(A|B)$ with $\Pr(B|A)$ or multiplying together probabilities of events, which are not independent, to find the probability of a compound event. The use of DNA evidence, results of lie detector tests, and the accuracy of eyewitness testimony have all been plagued by misuse of probability. See the References.

You can find a listing of references and suggestions for additional reading on the books's website, www.wiley.com/college/olinick

CHAPTER 11 Markov Processes

Often do the Spirits
Of great events
stride on before the events.
And in today already walks tomorrow . . .

—Samuel Taylor Coleridge

I. Markov Chains

A. Definitions

Markov chains have been and continue to be one of the most important and popular tools of mathematical model builders. This chapter presents some of the fundamental ideas of Markov chains and indicates some of their uses. More extended applications are presented in Chapters 12, 13, and 17. The necessary mathematical prerequisites for reading this chapter are the concepts of probability presented in Sections II (A–D) and IV of Chapter 10 and the elementary properties of matrix algebra discussed in Appendix II.

The fundamental principle underlying Markov processes is *the independence of the future from the past if the present is known*. Imagine an experiment that is repeated once each day for many days. If the probabilities of the outcomes of tomorrow's experiment depend only on the outcome of today's experiment and do not depend on the results of any previous experiments, then you are dealing with a Markov process.

In slightly different language, a finite Markov chain is a stochastic process with a finite number of states in which the probability of being in a particular state at the $(n+1)$st step depends only on the state occupied at the nth step; this dependence is the same at all steps. More formally, there is the following definition:

DEFINITION An experiment with a finite number of possible outcomes $S_1, S_2, \ldots, S_r$ is repeated a number of times. The sequence of outcomes is a *Markov chain* if there is a set of r^2 numbers $\{p_{ij}\}$ such that the conditional probability of outcome S_j on any experiment given outcome S_i on the previous experiment is p_{ij}—that is,

$$p_{ij} = \Pr(S_j \text{ on experiment } n+1 \mid S_i \text{ on experiment } n),\ 1 \le i, j \le r,\ n = 1, 2, \ldots$$

The outcomes $S_1, S_2, \ldots, S_r$ are called *states*, and the numbers p_{ij} (which depend only on i and j not on n) are called *transition probabilities*. The transition probabilities may be arranged in a matrix with r columns and r rows:

$$P = \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1r} \\ p_{21} & p_{22} & \cdots & p_{2r} \\ \cdots & & & \\ p_{r1} & p_{r1} & & p_{rr} \end{pmatrix} \tag{1}$$

which is called the *transition matrix* for the Markov chain.

Note that each entry of a transition matrix is nonnegative and that the sum of the numbers in each row of the matrix is 1.

EXAMPLE 1

A recent study focused on the relationship between the birth weights of English women and the birth weights of their daughters. The weights were split into three categories: low (below 6 pounds), average (between 6 and 8 pounds), and high (above 8 pounds). Among women whose own birth weights were low, 50% of the daughters had low birth weights, 45 percent had average weights, and 5% had high weights. Women with average birth weights had daughters with average weights half of the time, while the other half was split evenly between low and high categories. Women with high birth weights had female babies with high weights 40% of the time, with low and average weights each occurring 30% of the time.

Example 1 can be considered as a Markov chain with three states (low, average, high), corresponding to an "experiment" of choosing a woman at random and noting her birth weight. The transition matrix, easily derived from the verbal description, looks like this:

$$P = \begin{array}{c} \text{Mother} \\ \\ \textit{Low} \\ \textit{Average} \\ \textit{High} \end{array} \begin{array}{c} \text{Daughter} \\ \begin{array}{ccc} \textit{Low} & \textit{Average} & \textit{High} \end{array} \\ \begin{pmatrix} .5 & .45 & .05 \\ .25 & .5 & .25 \\ .3 & .3 & .4 \end{pmatrix} \end{array} \tag{2}$$

B. State Diagrams

Transition probabilities may be conveniently presented in a matrix. They may also be shown in what is called a *transition diagram* or *state diagram*. This is a graph with vertices corresponding to the states and a directed arc from vertex i to vertex j if the transition probability p_{ij} is positive. The numerical values of the transition probabilities are written alongside the arcs. A transition diagram for Example 1 is shown in Fig. 11.1.

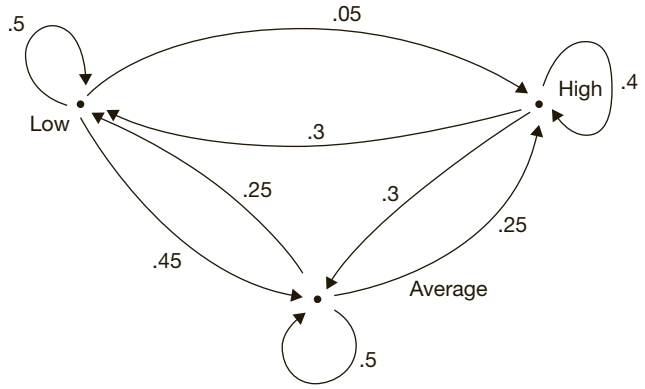


FIGURE 11.1 State diagram for Example 1 about birth weights.

Example 2

The Audio-Video Den, an electronics store in Milwaukee, has one item on special sale each day; it is either a television, a radio, or a stereo set. Stereos are never on sale two days in a row; if the store has a stereo as the special one day, it is equally likely to have a TV or radio on special the next day. If the special one day is a TV or a radio, there is an even chance of continuing the item the next day. If the special item is changed from a TV or radio, only one-third of the time will a stereo set be the special the next day.

Consider this as an example of Markov chain with states of TV, radio, and stereo. The transition matrix is

$$P = \begin{array}{c} \text{Today's Special} \\ \\ \\ \\ \\ \\ \\ \end{array} \begin{array}{c} \text{Tomorrow's Special} \\ \begin{array}{ccc} \textit{TV} & \textit{Radio} & \textit{Stereo} \end{array} \\ \begin{pmatrix} 1/2 & 1/3 & 1/6 \\ 1/3 & 1/2 & 1/6 \\ 1/2 & 1/2 & 0 \end{pmatrix} \end{array} \begin{array}{l} \\ \\ \textit{TV} \\ \textit{Radio} \\ \textit{Stereo} \\ \\ \\ \end{array} \tag{3}$$

and the state diagram is shown in Fig. 11.2.

The loops at the TV and radio vertices correspond to the fact that it is possible for either of these items to be repeated as the specials on consecutive days. The absence of a loop at the stereo vertex reflects the store's policy of never repeating the stereo as a special item on successive days.

Although the transition matrix is a powerful tool in analyzing Markov chains, the state diagram often reveals information about the process not immediately apparent from the matrix. For example, Fig. 11.2 indicates that no matter what item is the special today, it is possible for any one of the three items to be the special on the day after tomorrow.

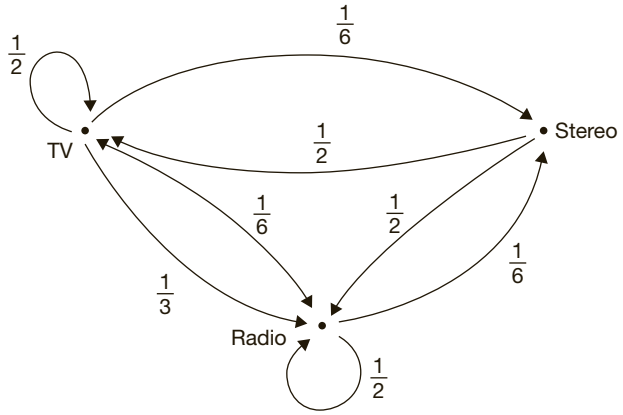


FIGURE 11.2 State diagram for Example 2 about the Audio-Video Den.

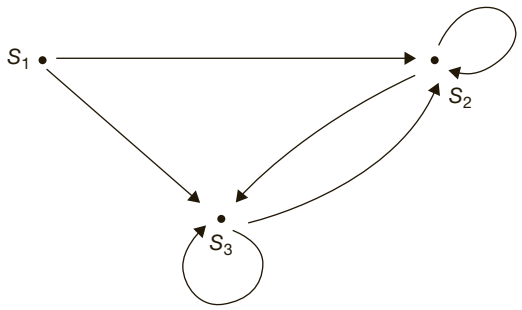


FIGURE 11.3

Fig. 11.3 presents a state diagram for another Markov chain. It is evident from this diagram that S_1 can be reached, at most, once. If the process is in state S_1 at the nth step, then it will be in state S_2 or S_3 at the $(n+1)$st step. From either of these states it is impossible to return to S_1.

The transition matrix for this Markov chain would have the form

$$P = \begin{array}{c} \\ S_1 \\ S_2 \\ S_3 \end{array} \begin{array}{c} \begin{array}{ccc} S_1 & S_2 & S_3 \end{array} \\ \begin{pmatrix} 0 & a & 1-a \\ 0 & b & 1-b \\ 0 & c & 1-c \end{pmatrix} \end{array} \tag{4}$$

for some positive numbers a, b, and c.

C. Tree Diagrams

Since Markov chains are particular examples of stochastic processes, they can be analyzed with the help of tree diagrams. For example, if a stereo is the special sale item today, we may be interested in the probability that a stereo will again be the special item 3 days from now. The tree diagram that enables us to answer this question is drawn in Fig. 11.4. Note that we have omitted the branches corresponding to a zero probability and have used the notation $S =$ Stereo, $R =$ Radio, $T =$ Television.

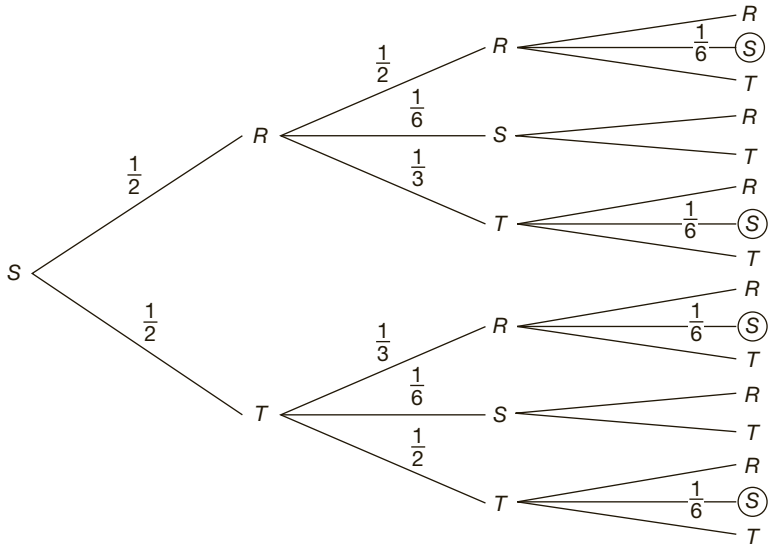


FIGURE 11.4 To find the probability that a stereo (S) is on sale 3 days from now, add the probabilities of each of the paths from the starting node S to the circled final S nodes.

From the tree diagram, we see that there are four distinct ways the stereo can be the special sale item 3 days from now: *RRS*, *RTS*, *TRS*, and *TTS*. Here *RRS* indicates that radios are the special items for tomorrow and the next day and a stereo the day after that. The desired probability is

$$\begin{aligned}&\Pr(RRS) + \Pr(RTS) + \Pr(TRS) + \Pr(TTS)\\ &= \left(\frac{1}{2}\right)\left(\frac{1}{2}\right)\left(\frac{1}{6}\right) + \left(\frac{1}{2}\right)\left(\frac{1}{3}\right)\left(\frac{1}{6}\right) + \left(\frac{1}{2}\right)\left(\frac{1}{3}\right)\left(\frac{1}{6}\right) + \left(\frac{1}{2}\right)\left(\frac{1}{2}\right)\left(\frac{1}{6}\right)\\ &= \frac{1}{24} + \frac{1}{36} + \frac{1}{36} + \frac{1}{24} = \frac{5}{36}\end{aligned}$$

In Section II, it will be shown how the same question can be more easily answered by use of the transition matrix and the technique of matrix multiplication.

D. Initial Probabilities

In order to describe the way a Markov chain develops, we need, in addition to the transition probabilities, a distribution of *initial probabilities* $(p_1, p_2, \ldots, p_r)$ where p_k is the probability that the outcome of the first experiment is S_k. In the study of birth weights for example, the first generation of women investigated had 25 percent of its members of low birth weight, 60 percent of average weight, and 15 percent of high weight. The initial probabilities would then be described by $(p_1, p_2, p_3) = (.25, .6, .15)$.

A Markov chain then operates in the following way. There is a stochastic process that moves from state to state in a sequence of steps. By means of the initial probability distribution, the process starts at one of the states S_k with probability p_k. If at any step it is in state S_i, then it moves to state S_j with probability p_{ij}. This probability is found in the

distribution of the ith row of the transition matrix. The entire process is completely described by the initial probability distribution (a $1 \times r$ vector) and the transition matrix.

E. An Absorbing Example

This section concludes with one more example of a Markov process.

Example 3

The Academic Personnel Committee at Lower Pine Cone College reviews the contracts of all faculty members each year. The rules of the college demand that a professor with tenure must be continued, but all other faculty members may be fired. If a faculty member is not fired, then he may be kept on for another year at the same rank, promoted to a tenured position, or promoted but not given tenure. However, if a professor was promoted and not given tenure the previous year, his next promotion must be to a tenured rank. Life is so pleasant at this college and the job market so dismal that no one leaves the school voluntarily: everyone is either eventually fired or given tenure.

This process may be viewed as a Markov chain. The states are the employment categories of an individual at the end of a particular year: fired (F), promoted to tenure rank (T), promoted but not with tenure (P), and retained at an untenured rank without promotion (R). The transition matrix then has the form

$$P = \begin{array}{l} \\ \text{Position at end of this year} \\ \\ \end{array} \begin{array}{c} \text{Position at end of next year} \\ \begin{array}{cccc} F & T & R & P \end{array} \\ \begin{array}{c} F \\ T \\ R \\ P \end{array} \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ a & b & c & d \\ e & f & g & 0 \end{pmatrix} \end{array} \tag{5}$$

where the numbers $a-g$ are nonnegative probabilities.

This Markov chain has a feature missing from the others we have examined. There are two states, F and T, which may be called *absorbing states*. Once the process enters one of these states, it never leaves.

II. Matrix Operations and Markov Chains

A. Stochastic Matrices

If all the entries of a matrix are nonnegative and the sum of the entries in each row is 1, then the matrix is called a *stochastic matrix*. Every transition matrix for a Markov chain is an example of a stochastic matrix. Stochastic matrices have an interesting property: whenever two of them are multiplied, the result is another stochastic matrix.

THEOREM 1 If A and B are stochastic matrices and the product AB is defined, then AB is a stochastic matrix.

Proof of Theorem 1 Suppose A is a $k \times m$ matrix and B is an $m \times n$ matrix. It is clear that all the entries of AB will be nonnegative (review Appendix II if matrix multiplication is not familiar to you). We must show that the sum of the elements in any row of AB is equal to 1. Now the sum of the entries in row i of AB is given by

$$\begin{aligned}
\sum_{j=1}^{n}(AB)_{ij} &= \sum_{j=1}^{n}\left(\sum_{s=1}^{m}A_{is}B_{sj}\right) && \text{(definition of matrix multiplication)} \\
&= \sum_{s=1}^{m}\left(\sum_{j=1}^{n}A_{is}B_{sj}\right) && \text{(reversing order of summation)} \\
&= \sum_{s=1}^{m}A_{is}\left(\sum_{j=1}^{n}B_{sj}\right) && \text{(factoring out } A_{is}\text{)} \\
&= \sum_{s=1}^{m}A_{is}(1) && \text{(sum of entries in any row of } B \text{ is 1)} \\
&= \sum_{s=1}^{m}A_{is} = 1 && \text{(sum of entries in any row of } A \text{ is 1).}
\end{aligned}$$

Since i was an arbitrarily chosen row number, the theorem is proved. ◇

Corollary If A is an $r \times r$ stochastic matrix, then so are $A^2, A^3, A^4, \ldots$.

As an example, consider the 2×2 stochastic matrix

$$A = \begin{pmatrix} .7 & .3 \\ .4 & .6 \end{pmatrix} \tag{6}$$

that has

$$A^2 = \begin{pmatrix} .61 & .39 \\ .52 & .48 \end{pmatrix}$$

and

$$A^3 = \begin{pmatrix} .583 & .417 \\ .556 & .444 \end{pmatrix}$$

As an exercise to be completed before reading any further, interpret the $r \times r$ stochastic matrix as the transition matrix of a Markov chain and prove the corollary directly using probabilistic considerations only. ◇

B. Probability Distribution after *n* Steps

One of the most important questions about Markov chains is this: if the process begins in state S_i, what is the probability that after n steps it will be in state S_j? Denote this probability by $p_{ij}^{(n)}$. If we are interested in this problem for all possible starting states S_i, and terminating states S_j, the probabilities may be represented in a matrix,

$$P^{(n)} = \begin{pmatrix} p_{11}^{(n)} & p_{12}^{(n)} & \cdots & p_{1r}^{(n)} \\ p_{21}^{(n)} & p_{22}^{(n)} & \cdots & p_{2r}^{(n)} \\ \cdots & & & \\ p_{r1}^{(n)} & p_{r2}^{(n)} & \cdots & p_{rr}^{(n)} \end{pmatrix} \tag{7}$$

or a tree diagram (Fig. 11.5).

In Section I.C, we saw how such a problem can be solved through the use of tree diagrams. Tree diagrams are convenient, however, only when n is a relatively small number. If n is large, the number of branches in the corresponding tree diagram is too great to draw easily. In this section, we will see how the question can be answered using matrix multiplication.

If $n = 1$, then $p_{ij}^{(n)} = p_{ij}^{(1)}$ is the probability of moving from state S_i to state S_j in one step. By definition of a Markov process, this is exactly the transition probability p_{ij}. Thus, $p_{ij}^{(1)} = p_{ij}$ for all i and j, and $P^{(1)} = P$.

Determine next $p_{ij}^{(2)}$. The computation is facilitated by examining the relevant portions of the tree diagram in Fig. 11.5.

Note that

$$p_{ij}^{(2)} = p_{i1}p_{1j} + p_{i2}p_{2j} + \cdots + p_{ir}p_{rj} = \sum_{k=1}^{r} p_{ik}p_{kj}\text{r}$$

The definition of matrix multiplication, however, asserts that this number is just the ijth entry of P^2. In other words, $p_{ij}^{(2)} = (P^2)_{ij}$ so that $P^{(2)} = P^2$. The matrix that gives the probability distribution after two steps is the square of the transition matrix.

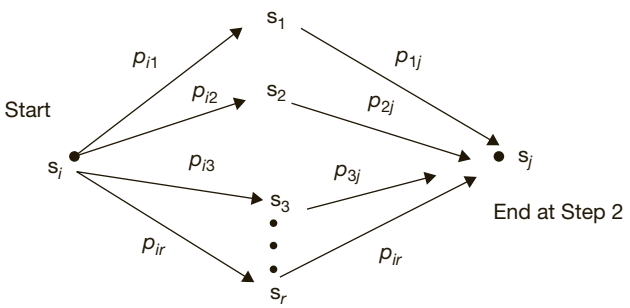


FIGURE 11.5

With these first two steps completed, the general result is easy to guess: the matrix of probability distributions after n steps is the nth power of the original transition matrix. The proof follows by induction on n. For later reference, we list the result as

THEOREM 2 For a Markov chain, $P^{(n)} = P^n$.

As noted in Section I, a Markov chain is determined by the transition matrix and a distribution of initial probabilities. Suppose the initial states are given by probabilities $p_i^{(0)}$, $i = 1, 2, \ldots, r$. Write the initial probability distribution as a row vector

$$\mathbf{p}^{(0)} = (p_1^{(0)}, p_2^{(0)}, \ldots p_r^{(0)})$$

If $p_j^{(n)}$ denotes the probability of being in state S_j after n steps, the vector of these probabilities is

$$\mathbf{p}^{(n)} = (p_1^{(n)}, p_2^{(n)}, \ldots p_r^{(n)})$$

Note that the sum of the entries in each of these vectors is 1—that is, $\mathbf{p}^{(n)}$ is a stochastic matrix of dimension $1 \times r$.

We may now obtain the critical relation between these probability row vectors and the transition matrix. Suppose we wish to compute $p_j^{(n)}$. The tree diagram of Fig. 11.6 tells what to do.

We have

$$\begin{aligned} p_j^{(n)} &= p_1^{(n-1)} p_{1j} + p_2^{(n-1)} p_{2j} + \cdots + p_r^{(n-1)} p_{rj} \\ &= \sum_{k=1}^{r} p_k^{(n-1)} p_{kj} \end{aligned}$$

Now this last sum is simply the product of the row vector $\mathbf{p}^{(n-1)}$ and the jth column of the transition matrix P. Thus, the components of $\mathbf{p}^{(n)}$ are obtained

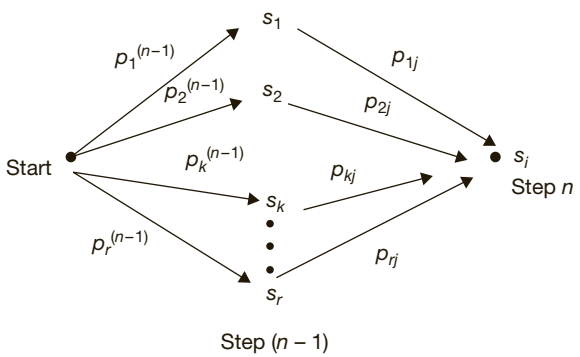


FIGURE 11.6 Determining $p_j^{(n)}$, the probability of being in state Sj after $(n-1)$ steps and the transition matrix.

by multiplying $\mathbf{p}^{(n-1)}$ by the appropriate column of P. This gives the important relationship

$$\mathbf{p}^{(n)} = \mathbf{p}^{(n-1)}P \tag{8}$$

From this relationship, we obtain

$$\mathbf{p}^{(1)} = \mathbf{p}^{(0)}P \tag{9}$$

$$\mathbf{p}^{(2)} = \mathbf{p}^{(1)}P = (\mathbf{p}^{(0)}P)P = \mathbf{p}^{(0)}P^2 \tag{10}$$

$$\mathbf{p}^{(3)} = \mathbf{p}^{(2)}P = (\mathbf{p}^{(0)}\mathbf{P}^2)P = \mathbf{p}^{(0)}P^3 \tag{11}$$

A straightforward induction argument establishes the general result stated in the next theorem. ◇

THEOREM 3 For any Markov chain, $\mathbf{p}^{(n)} = \mathbf{p}^{(0)}P^n$.

Review of notation

p_{ij} = probability of moving from state S_i to state S_j in one step.

$P = r \times r$ transition matrix whose entries are p_{ij}.

$p_{ij}^{(n)}$ = probability of moving from S_i to S_j in exactly n steps.

$P^{(n)} = r \times r$ matrix whose entries are $p_{ij}^{(n)}$.

$p_j^{(n)}$ = probability of being in state S_j after n steps.

$p^{(n)} = 1 \times r$ vector whose entries are $p_j^{(n)}$.

Main results:

1. $P^{(n)} = P^n$.

2. $p^{(n)} = p^{(0)}P^n$. ◇

C. Applications

The result of Theorem 3 can be used to answer the question about the Audio-Video Den stated in Section I.C. In this Markov chain, S_1 = television, S_2 = radio, and S_3 = stereo. The question concerned the probability that a stereo will be the special sale item 3 days hence if it is the sale item today. Taking today as the 0th step of the process, we have $\mathbf{p}^{(0)} = (0, 0, 1)$. By Theorem 3, the probability distribution after three steps is $\mathbf{p}^{(3)} = (0, 0, 1)P^3$. Now the

row vector $(0, 0, 1)P^3$. will simply be the third row of P^3. Carrying out the matrix multiplication gives

$$P^2 = \begin{pmatrix} 16/36 & 15/36 & 5/36 \\ 15/36 & 16/36 & 5/36 \\ 15/36 & 15/36 & 6/36 \end{pmatrix} \tag{12}$$

and

$$P^3 = \begin{pmatrix} 93/216 & 92/216 & 31/216 \\ 92/216 & 93/216 & 31/216 \\ 93/216 & 93/216 & 30/216 \end{pmatrix} \tag{13}$$

so that $\mathbf{p}^{(3)} = (\frac{93}{216}, \frac{93}{216}, \frac{30}{216})$, and the probability of being in state S_3 after three steps is the final entry of this vector, $\frac{30}{216} = \frac{5}{36}$.

Once the matrix P^3 is determined, any question about the nature of this Markov chain during its first three steps can be answered.

As another illustration of the use of Theorem 3, consider the birth weight model of Example 1. Suppose the initial generation of mothers surveyed contained 25% low birth weight women, 60% average weight, and 15% high weight. What would the distribution look like for the generation of their great-great-granddaughters? Since $S_l = \text{low}$, $S_2 = \text{average}$, $S_3 = \text{high}$, we have $\mathbf{p}^{(0)} = (.25, .6, .15)$ and we need to find $\mathbf{p}^{(4)} = \mathbf{p}^{(0)}P^4$, where P is the transition matrix (2). In this case, the fourth power of the transition matrix is

$$P^4 = \begin{pmatrix} .347969 & .442762 & .209269 \\ .346813 & .438719 & .214469 \\ .348113 & .438863 & .213025 \end{pmatrix}$$

and the required probability vector is

$$\mathbf{p}^{(4)} = (.347297, .439751, .212952).$$

The conclusion then is that about 35% of the great-great-granddaughters will have low birth weights, 44% average weights, and 21% high weights.

As a final application, consider the tenure model of Example 3. Over the past 10 years, the Academic Personnel Committee has consistently followed a pattern of promotion described by the transition probabilities

$$a = .1,\ b = .01, c = .67, d = .22, e = .05, f = .45, g = .5.$$

so that the transition matrix has the form

$$\begin{array}{c} \\ F \\ T \\ R \\ P \end{array} \begin{array}{c} \begin{array}{cccc} F & T & R & P \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ .1 & .01 & .67 & .22 \\ .05 & .45 & .50 & 0 \end{pmatrix} \end{array} \tag{14}$$

Suppose there are now 100 members of the faculty, distributed 30, 40, 30 in the states T, R, P, respectively. If the personnel committee continues its past policies for another 5 years, let us determine the status of this group of professors at the end of that period.

For simplicity, assume that no one on the faculty leaves the system through retirement, death, or some voluntary action. Let $\mathbf{p}^{(0)} = (0, .3, .4, .3)$ represent the initial distribution. We seek $\mathbf{p}^{(5)} = \mathbf{p}^{(0)}P^5$. Matrix multiplication gives

$$P^5 = \begin{array}{c} \\ F \\ T \\ R \\ P \end{array} \begin{array}{c} \begin{array}{cccc} F & \mathrm{T} & R & P \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ .33 & .3 & .29 & .08 \\ .2 & .57 & .18 & 05 \end{pmatrix} \end{array}$$

where the entries have been rounded off to two decimal places. Thus, the distribution after 5 years is

$$\mathbf{p}^{(5)} = (.19, .59, .17, .05).$$

The prediction of this model is that 19 of the untenured faculty members will be fired, 29 will have advanced to tenured rank so that the faculty will have 59 tenured members, 17 will have been retained but without promotion, and 5 will be given promotions without tenure.

Although this model cannot predict what happens to a particular untenured professor during the 5 years, it is still a quite useful tool for planning and decision making. The personnel committee can use the model to predict the cumulative effects of its past policies if they are continued into the future or to assess the effects of proposed changes in the transition probabilities.

The next two sections, III and IV, present a more detailed mathematical treatment of two particular types of Markov processes: regular chains and absorbing chains.

III. Regular Markov Chains

A. Definitions

If you glance back at the matrix P^3 of Eq. (13),

$$\begin{array}{c} \\ TV \\ Radio \\ Stereo \end{array} \begin{array}{c} \begin{array}{ccc} TV & Radio & Stereo \end{array} \\ \begin{pmatrix} 93/216 & 92/216 & 31/216 \\ 92/216 & 93/216 & 31/216 \\ 93/216 & 93/216 & 30/216 \end{pmatrix} \end{array} \tag{13}$$

Table 11.1

$\mathbf{p}^{(0)}$	$\mathbf{p}^{(4)}$
(.25, .6, .15)	(.347297, .439751, .212952)
$\left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$	(.347631, .440115, .212254)
(1, 0, 0)	(.347969, .442762, .209269)
(.1, 8, .1)	(.347058, .439138, .213804)
(0, 0, 1)	(.348113, .438863, .213025)

it may surprise you to see that the rows of the matrix are almost identical. In the discussion of the Audio-Video Den, for which the matrix was derived, we saw that the probability of being in the "stereo state" after three steps, if the process began in the stereo state, is the third component of the third row of P^3—that is, 30/216. We can also read from the matrix P^3 that the probability of being in the stereo state after three steps, if the starting state is "TV" or "radio" is practically the same: 31/216. It appears, then, that the long-range probability that a stereo is the special sale item may, in fact, be independent of what item is the special when the process starts. Inspection of the first and second columns of P^3 indicates that this may be true for televisions and radios as well.

As another illustration, consider the distribution of birth weights of great-great-granddaughters for different initial distributions of birth weights of the original generation of mothers. Some of these are presented in Table 11.1.

Note that despite wide variations in the choice of an initial distribution, the probability distribution after four steps is always very close to (.35, .44, .21).

The type of behavior shown by these two examples occurs whenever the underlying Markov chain process possesses a regularity property.

DEFINITION A Markov process is a *regular chain* if some power of the transition matrix has only positive entries.

In particular, the Markov process is regular if all entries in the transition matrix $P = P^1$ are positive. Thus, the birth weight model of Example 1 is a regular chain. The matrix of transition probabilities for the Audio-Video Den illustration (Example 2) contains a 0 entry, but its second power, P^2, has all entries positive (Eq. (12)). Thus, this is a regular chain also.

If the transition matrix of a Markov process is an identity matrix, then so is every power of that matrix and the underlying chain is not regular. The tenure model (Example 3) is not a regular chain either; every power of the transition matrix will have its first two rows identical to the first two rows of P and hence will contain entries equal to 0.

A Markov process is a regular one if there is some positive integer n, so that the process may be in any one of the possible states n steps after starting, regardless of the initial state. The smallest n for which this is possible is the smallest positive integer n for which P^n has no zero entries.

If P is the transition matrix of a regular Markov chain, then it turns out that the powers of P approach a matrix W, all of whose rows are the same. If $\mathbf{w}$ denotes the row vector formed from any of the rows of W, then it also happens that $\mathbf{w}P = \mathbf{w}$. There results will be formalized, proved, and applied in the next few pages.

DEFINITION A vector $\mathbf{w}$ is a fixed-point *vector* of the matrix P if $\mathbf{w}P = \mathbf{w}$. A Markov chain is said to be in *equilibrium* if the probability distribution at some step is given by a fixed-point vector of the transition matrix.

Note that if $\mathbf{w}P = \mathbf{w}$, then $\mathbf{w}P^2 = (\mathbf{w}P)P = \mathbf{w}P = \mathbf{w}$, and, in general, $\mathbf{w}P^n = \mathbf{w}$.

Example 4

The zero vector $0 = (0, 0, \ldots, 0)$ is a fixed-point vector of every transition matrix.

Example 5

If P is the transition matrix

$$P = \begin{pmatrix} .7 & .3 \\ .4 & .6 \end{pmatrix}$$

and $\mathbf{w}$ is the vector $\mathbf{w} = (\frac{4}{7}, \frac{3}{7})$, it is easy to check whether $\mathbf{w}P = \mathbf{w}$.

Example 6

If P is any 2×2 matrix,

$$P = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \text{ with } b + c \neq 0\,, \tag{15}$$

then the vector

$$W = \left(\frac{c}{b+c}, \frac{b}{b+c}\right)$$

is a fixed-point vector of P.

B. First Basic Theorem

THEOREM 4 (FIRST BASIC THEOREM FOR REGULAR MARKOV CHAINS) If P is the transition matrix for a regular Markov chain, then

(a) the powers P^n approach a stochastic matrix W;

(b) each row of W is the same vector $\mathbf{w} = (w_1, w_2, \ldots, w_r)$; and

(c) the components of $\mathbf{w}$ are positive.

In order to prove this theorem, we first establish a helpful lemma.

LEMMA Suppose P is an $r \times r$ transition matrix having no zero entries and let q be the smallest entry of P. Let $\mathbf{x}$ be an $r \times 1$ column vector, having largest component M_0 and smallest component m_0. Let M_1 and m_1 be the largest and smallest components of the vector $P\mathbf{x}$. Then

1. $M_1 \leq M_0$

2. $m_1 \geq m_0$

3. $M_1 - m_1 \leq (1 - 2q)(M_0 - m_0)$ ◇

Example 7

If P is the transition matrix (2) of the birth weight model—that is,

$$P = \begin{array}{c} \\ \textit{Low} \\ \textit{Average} \\ \textit{High} \end{array} \begin{array}{c} \begin{array}{ccc} \textit{Low} & \textit{Average} & \textit{High} \end{array} \\ \begin{pmatrix} .5 & .45 & .05 \\ .25 & .5 & .25 \\ .3 & .3 & .4 \end{pmatrix} \end{array}$$

then we have $q = .05$. If $\mathbf{x}$ is the vector

$$\mathbf{x} = \begin{pmatrix} .2 \\ .3 \\ .5 \end{pmatrix}$$

then $M_0 = .5$ and $m_0 = .2$.

According to the lemma, the largest component of $P\mathbf{x}$ will be at most .5, the smallest component will be at least .2, and the difference between these components, $M_1 - m_1$, will be at most

$$(1 - 2[.05])(.5 - .2) = (.9)(.3) = .27$$

This claim may be checked by carrying out the indicated matrix multiplication. We obtain

$$P\mathbf{x} = \begin{pmatrix} .26 \\ .325 \\ .35 \end{pmatrix}$$

so that $M_1 = .35 < .5 = M_0$ and $m_1 = .26 > .2 = m_0$. The difference satisfies $M_1 - m_1 = .35 - .26 = .09 < .27$.

Before proceeding to a proof of the lemma, note that if all entries of a transition matrix P are positive, then we must have $0 < q \leq \frac{1}{2}$ since the sum of the entries in each row is 1. Thus, $0 \leq 1 - 2q < 1$. In particular, the lemma asserts that $(M_1 - m_1) < (M_0 - m_0)$. The effect of applying the matrix P to the vector $\mathbf{x}$ is to produce a vector $P\mathbf{x}$ whose components are more nearly equal than the components of $\mathbf{x}$.

Proof of Lemma The ith component of $P\mathbf{x}$ is the product of the ith row of the matrix P and the vector $\mathbf{x}$—that is,

$$(P\mathbf{x})_i = \begin{pmatrix} p_{i1} & p_{i2} & \dots & p_{ir} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \dots \\ x_r \end{pmatrix} = p_{i1}\, x_1 + p_{i2}\, x_2 + \cdots + p_{ir}\, x_r \tag{16}$$

Since $p_{i1} + p_{i2} + \cdots + p_{ir} = 1$, the ith component of $P\mathbf{x}$ may be regarded as the expected value of a gamble whose outcomes are the components of $\mathbf{x}$ that occur with probabilities given by the entries of the ith row of P. Considering each of the rows of P as a different gamble, the number M_1 measures the expected value of the most favorable gamble. We shall concentrate on this particular gamble.

If the outcomes of the gamble are changed so that one of them is m_0 and all the rest are M_0, then the new gamble will have an expected value at least as large as the original gamble. Now the largest possible expected value for such a gamble occurs if the smallest outcome, m_0, occurs with the smallest probability, q. In this case, the expected value is $qm_0 + (1 - q)M_0$. Thus, we have

$$M_1 \leq qm_0 + (1 - q)M_0 \leq qM_0 + (1 - q)M_0 = M_0 \tag{17}$$

establishing the inequality $M_1 \leq M_0$.

A similar argument, based on considering the least favorable gamble, shows that

$$m_1 \geq qM_0 + (1 - q)m_0 \geq qm_0 + (1 - q)m_0 = m_0 \tag{18}$$

so that $m_1 \geq m_0$.

Multiply the first inequality of (18) by (–1) and add to the first inequality of (17) to obtain

$$M_1 - m_1 \leq q(m_0 - M_0) + (1 - q)(M_0 - m_0) = (1 - 2q)(M_0 - m_0)$$

completing the proof of the lemma. ◇

Proof of Theorem We deal first with the case when all entries of P are positive. Suppose $\mathbf{x}$ is a column vector. Let q be the smallest entry in P and let M_n and m_n denote the largest and smallest components of $P^n\mathbf{x}$, respectively.

Since $P^n\mathbf{x} = P(P^{n-1}\mathbf{x})$, repeated applications of the lemma give

$$M_1 \geq M_2 \geq M_3 \geq \cdots \tag{19}$$

$$m_1 \leq m_2 \leq m_3 \leq \cdots \tag{20}$$

and

$$M_n - m_n \leq (1 - 2\mathrm{q})(M_{n-1} - m_{n-1}) \tag{21}$$

so that

$$M_n - m_n \leq (1 - 2q)^n(M_0 - m_0) \tag{22}$$

Since $1 - 2q$ is less than 1, $(1 - 2q)^n$ tends to 0 as n gets large. Thus, the difference $M_n - m_n$ also goes to zero. This implies that M_n and m_n approach a common limit and $P^n\mathbf{x}$, tends to a vector all of whose components are the same.

This common value will lie between m_n and M_n, *for* all n. Since we have $0 < m_0 \leq M_0 < 1$, the common value is a strictly positive number less than 1.

Now let $\mathbf{x}$ be the column vector with kth component equal to 1 and all other components 0. Then $P^n\mathbf{x}$ is simply the kth column of P^n. We have shown that the kth column of P^n tends to a vector with all components equal. Denote the common value of the components by w_k. Thus, P^n tends to a matrix W with all rows the same vector $\mathbf{w} = (w_1, w_2, \ldots, w_r)$

Since the sum of the entries in each row of P^n is always 1, regardless of n, the same must be true of the limit W (because the limit of a sum is the sum of the limits). Thus, W is a stochastic matrix. This establishes the theorem in the case that all entries of P are positive.

In the general case of a regular Markov chain, the transition matrix P may have some zero entries. Since the Markov chain is regular, some power P^N of P has all positive entries. If q^* is the smallest entry of P^N, then the first part of the proof shows that

$$(M_{N+1} - m_{N+1}) \leq (1 - 2q*)(M_N - m_N)$$

The sequence $\{d_n\}$ where $d_n = M_n - m_n$ is then a nonincreasing sequence with a subsequence $\{d_{n+N}\}$ tending to 0. This forces the entire sequence $\{d_n\}$ to have limit 0. The rest of the proof is the same as the proof of the special case. ◇

C. Second Basic Theorem

The next important theorem shows an easy way to find the limiting matrix W of a regular Markov chain.

THEOREM 5 (SECOND BASIC THEOREM FOR REGULAR MARKOV CHAINS) If P is the transition matrix for a regular Markov chain and W and $\mathbf{w}$ are the matrix and vector promised by Theorem 4, then

(a) For any stochastic row vector $\mathbf{p}$, $\mathbf{p}P^n$ approaches $\mathbf{w}$

(b) The vector $\mathbf{w}$ is the unique fixed-point stochastic vector of P

Proof of Theorem 5 Since $P^n \to W$, we have $\mathbf{p}P^n \to \mathbf{p}W$. Every entry in the kth column of W is w_k so the kth component of $\mathbf{p}W$ is equal to w_k multiplied by the sum of the entries of $\mathbf{p}$. Since $\mathbf{p}$ is a stochastic vector, that sum is 1. Thus, the kth component of $\mathbf{p}W$ is w_k. In other words, $\mathbf{p}W = \mathbf{w}$. This proves (a).

To prove (b), note that the powers of P approach W so that $P^{n+1} = P^n P$ approaches W also. But $P^n P$ also approaches WP. Thus, $W = WP$.

Each row of the matrix equation $W = WP$ simply asserts that $\mathbf{w} = \mathbf{w}P$ so that $\mathbf{w}$ is a fixed-point vector of P. By Theorem 4, $\mathbf{w}$ is a stochastic vector. All that is left to show is the uniqueness of $\mathbf{w}$. Accordingly, suppose v is any stochastic fixed-point vector of P. The $\mathbf{v}P^n$ approaches $\mathbf{w}$. But $\mathbf{v}$ is a fixed-point vector for P, so that $\mathbf{v}P^n = \mathbf{v}$ for all n. Thus, $\mathbf{v}$ approaches $\mathbf{w}$. But $\mathbf{v}$ is a constant vector. Hence, $\mathbf{v} = \mathbf{w}$. This completes the proof of Theorem 5. ◇

If P is the transition matrix of a regular Markov chain and $\mathbf{p}^{(0)}$ is the initial probability distribution, then Theorem 5 implies that $\mathbf{p}^{(0)}P^n$ approaches $\mathbf{w}$, the unique fixed-point stochastic vector of P, regardless of the particular numerical values of the entries of $\mathbf{p}^{(0)}$. We have already shown, however, that $\mathbf{p}^{(0)}P^n = \mathbf{p}^{(n)}$, the probability distribution after n steps. Thus, $\mathbf{p}^{(n)}$ approaches $\mathbf{w}$. In other words, no matter what the initial probabilities are for a regular Markov chain, after a large number of steps the probability that the process is in a particular state S_k will be very nearly W_k: a regular Markov chain approaches equilibrium.

To illustrate the approach to equilibrium, consider the transition matrix of Example 5:

$$P = \begin{pmatrix} .7 & .3 \\ .4 & .6 \end{pmatrix}$$

A fixed-point vector $\mathbf{w} = (w_1, w_2)$ for P must satisfy $\mathbf{w}P = \mathbf{w}$—that is,

$$(w_1, w_2)\, P = (w_1, w_2) \tag{23}$$

which is equivalent to

$$\begin{aligned} .7w_1 + .4w_2 &= w_1 \\ .3w_1 + .6w_2 &= w_2 \end{aligned} \tag{24}$$

or

$$\begin{aligned} -.3w_1 + .4w_2 &= 0 \\ .3w_1 - .4w_2 &= 0 \end{aligned} \tag{25}$$

and this system is satisfied by any pair of numbers, w_1 and w_2 such that $w_2 = \frac{3}{4}w_1$. Since a stochastic vector must have $w_1 + w_2 = 1$, we have

$$w_1 + \frac{3}{4}w_1 = 1$$

This gives $w_1 = \frac{4}{7}$ and $w_2 = \frac{3}{7}$. The unique fixed-point stochastic vector for the transition matrix P is

$$w = \left(\frac{4}{7}, \frac{3}{7}\right) = (.571428, .428572)$$

The first few powers of P are given by

$$P^2 = \begin{pmatrix} .61 & .39 \\ .52 & .48 \end{pmatrix} \qquad P^3 = \begin{pmatrix} .583 & .417 \\ .556 & .444 \end{pmatrix}$$

$$P^4 = \begin{pmatrix} .5749 & .4251 \\ .5668 & .4332 \end{pmatrix} \qquad P^5 = \begin{pmatrix} .57247 & .42753 \\ .57004 & .42996 \end{pmatrix}$$

$$P^6 = \begin{pmatrix} .571741 & .428259 \\ .571012 & .428988 \end{pmatrix} \quad P^7 = \begin{pmatrix} .571522 & .428478 \\ .571304 & .428696 \end{pmatrix}$$

$$P^8 = \begin{pmatrix} .571457 & .428543 \\ .571391 & .428609 \end{pmatrix} \quad P^9 = \begin{pmatrix} .571437 & .428563 \\ .571417 & .428583 \end{pmatrix}$$

and we see that P^n does approach W. If the Markov chain starts with initial probability vector $\mathbf{p}^{(0)} = (.9, .1)$, then the distributions after the first 10 steps are

$$\mathbf{p}^{(1)} = (.67, .33)$$
$$\mathbf{p}^{(2)} = (.601, .399)$$
$$\mathbf{p}^{(3)} = (.5803, .4197)$$
$$\mathbf{p}^{(4)} = (.57409, .42591)$$
$$\mathbf{p}^{(5)} = (.572227, .427773)$$
$$\mathbf{p}^{(6)} = (.571668, .428332)$$
$$\mathbf{p}^{(7)} = (.571500, .428500)$$
$$\mathbf{p}^{(8)} = (.57145, .42855)$$
$$\mathbf{p}^{(9)} = (.571435, .428565)$$
$$\mathbf{p}^{(10)} = (.57143, .428569)$$

D. Markov Processes as Discrete Dynamical Systems

We can gain a different insight into the behavior of a simple Markov process by regarding it as a linked system of difference equations. If we have a two-state Markov process, then we may write the transition matrix P as

$$P = \begin{pmatrix} p & 1-p \\ q & 1-q \end{pmatrix}$$

where the probabilities p and q lie between 0 and 1.

If we let A_n be the probability that the process is in the first state at step n, and B_n the probability that it occupies the second state at this step, the transition matrix gives us a discrete system of equations

$$\begin{aligned} A_{n+1} &= p\,A_n + q\,B_n \\ B_{n+1} &= (1-p)A_n + (1-q)B_n \end{aligned}$$

Since $A_n + B_n = 1$, we may rewrite the first of these equations as

$$A_{n+1} = pA_n + q(1 - A_n) = (p - q)\,A_n + q$$

From Chapter 1, we know that the solution of this difference equation is

$$A_n = (p-q)^n A_0 + q\frac{1-(p-q)^{n-1}}{1-(p-q)}$$

The probabilities p and q are between 0 and 1, so that $|p - q| < 1$, and the powers of $(p - q)$ will converge to 0. Thus,

$$\lim_{n\to\infty} A_n = \frac{q}{1-p+q}$$

Hence, in the long term, the process will occupy the first state with probability $\dfrac{q}{1-p+q}$ and the second state with probability $1 - \dfrac{q}{1-p+q} = \dfrac{q}{1-p}$.

In Example 5, we have $p = 7/10$ and $q = 4/10$ so this approach also predicts that the process will, in the long run, be in the first state with probability $\dfrac{\frac{7}{10}}{1-\frac{7}{10}+\frac{4}{10}} = \dfrac{4}{7}$.

E. Applications

This section on regular Markov chains concludes with some illustrations of how Theorems 4 and 5 can be applied to certain mathematical models.

Example 8

A study of "brand loyalty" in the antidandruff shampoo market showed that 70% of consumers who bought Flakes No More would buy it again when it was time to repurchase shampoo, and 30% would switch to the other available brand, Head, Neck, and Shoulders. The study also showed that 40% of Head, Neck, and Shoulders users would switch to Flakes No More while 60% would continue with the brand. In the long run, how much of the market can Flakes No More capture?

Solution

This is a regular Markov chain with two states, customer choosing Flakes No More (F) or Head, Neck, and Shoulders (H). The transition matrix is

$$P = \begin{array}{cc} & \begin{array}{cc} \text{Next purchase} & \\ F & H \end{array} \\ \begin{array}{c} \text{This} \\ \text{purchase} \end{array} & \begin{array}{c} F \\ H \end{array} \begin{pmatrix} .7 & .3 \\ .4 & .6 \end{pmatrix} \end{array}$$

Since the unique fixed-point stochastic vector for P is $\mathbf{w} = \left(\frac{4}{7}, \frac{3}{7}\right)$, in the long run $\frac{4}{7}$ of the population will be using Flakes No More and $\frac{3}{7}$ will be using Head, Neck, and Shoulders.

Example 9

According to the birthrate model of Example 1, what are the long-term distributions of birth weights among female babies?

Solution

We need to compute the fixed-point stochastic vector for the transition matrix P of Eq. (2). This leads to the matrix equation

$$(w_1, w_2, w_3)\begin{pmatrix} .5 & .45 & .05 \\ .25 & .5 & .25 \\ .3 & .3 & .4 \end{pmatrix} = (w_1, w_2, w_3)$$

which becomes the system

$$\begin{aligned} .5w_1 + .25w_2 + .3w_3 &= w_1 \\ .45w_1 + .5w_2 + .3w_3 &= w_2 \\ .05w_1 + .25w_2 + .4w_3 &= w_3 \end{aligned}$$

and this is equivalent to the homogeneous system

$$\begin{aligned} -.5w_1 + .25w_2 + .3w_3 &= 0 \\ .45w_1 - .5w_2 + .3w_3 &= 0 \\ .05w_1 + .25w_2 - .6w_3 &= 0 \end{aligned}$$

A solution to this system is any triple of numbers w_1, w_2, w_3 such that $w_1 = \frac{90}{55} w_3$ and $w_2 = \frac{114}{55} w_3$. To find a stochastic vector, impose the extra condition that $w_1 + w_2 + w_3 = 1$. This gives $w_3 = \frac{55}{259}$ so that the fixed-point stochastic vector is

$$\mathbf{w} = \left(\frac{90}{259}, \frac{111}{259}, \frac{55}{259}\right) = (.34749, .440155, .212355)$$

Hence, about 35% of the births will be in the low range, 44% in the average range, and 21% in the high range.

Chapter 12, which you may wish to read at this time, gives an extended discussion of a mathematical model in anthropology that makes critical use of a regular Markov chain to investigate certain questions about cultural stability.

IV. Absorbing Markov Chains

This section offers a detailed look at another important class of Markov chains frequently used in mathematical models of social and biological phenomena.

A. Definitions and Questions

A state S_k in a Markov chain is called an *absorbing state* if it is impossible to leave it—that is, the transition probabilities satisfy

$$\begin{cases} p_{kk} = 1 \\ p_{kj} = 0 \quad \text{if} \quad j \neq k \end{cases}$$

A Markov chain is an *absorbing chain* if it has at least one absorbing state and from every state it is possible to reach some absorbing state in a finite number of steps. If a state is not an absorbing state, it is called a *transient state*.

The Markov chain describing the personnel policies of Lower Pine Cone College (Example 3) is an absorbing chain with two absorbing states (F and T) and two transient states (R and P). It is possible to reach either of the absorbing states from either of the transient states in a single step.

The state diagram of Fig. 11.7 represents an absorbing Markov with the absorbing states (S_1, S_4, and S_6). From state S_2, the process can move to S_4 in two steps. From state S_3, the process may move to S_4 in one step. From state S_5, the process can move to S_4 (two steps) or S_6 (one step).

Some of the important questions about absorbing Markov chains follow:

1. Will the process eventually reach an absorbing state?
2. What is the average number of times we can expect the process to be in one transient state if it starts in another (or the same) transient state?

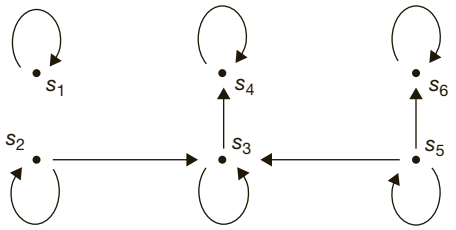


FIGURE 11.7 State diagram illustrating a Markov chain with three absorbing states (S_1, S_4, and S_6).

3. What is the average number of steps before the process enters an absorbing state?
4. What is the probability that the process will be absorbed in a particular state if it starts in a given transient state?

Our procedure in this section will be to introduce first the notation by which the answers to these questions may be presented, state the answers to the questions, illustrate with a few examples, and then give the proofs of the relevant theorems.

B. Notation and Answers

There is a standard way of representing the transition matrix of an absorbing chain: list the absorbing states first and then the transient states. For the Lower Pine Cone College example, the transition matrix is already in standard form:

$$P = \begin{array}{c} \\ F \\ T \\ R \\ P \end{array} \begin{array}{c} \begin{array}{cccc} F & T & R & P \end{array} \\ \left(\begin{array}{cc|cc} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ \hline a & b & c & d \\ e & f & g & 0 \end{array}\right) \end{array} \tag{26}$$

A standard form for the transition matrix of the chain whose state diagram is given by Fig. 11.7 is

$$P = \begin{array}{c} \\ S_1 \\ S_4 \\ S_6 \\ S_2 \\ S_3 \\ S_5 \end{array} \begin{array}{c} \begin{array}{ccccccc} S_1 & S_4 & S_6 & S_2 & S_3 & S_5 \end{array} \\ \left(\begin{array}{ccc|ccc} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & a & 1-a & 0 \\ 0 & b & 0 & 0 & 1-b & 0 \\ 0 & 0 & c & 0 & d & 1-c-d \end{array}\right) \end{array} \tag{27}$$

If an absorbing Markov process with r states has k absorbing states, then the transition matrix has the standard form:

$$\begin{array}{c} \\ \text{Absorbing States} \\ \text{Transient States} \end{array} \begin{array}{c} \begin{array}{cc} \text{Absorbing} & \text{Transient} \\ \text{States} & \text{States} \end{array} \\ \left(\begin{array}{c|c} I & 0 \\ \hline R & Q \end{array}\right) \end{array}$$

where
I is the $k \times k$ identity matrix
0 is the $k \times (r-k)$ zero matrix
R is a $(r-k) \times k$ matrix
Q is a $(r-k) \times (r-k)$ matrix

If I is the $(r-k) \times (r-k)$ identity matrix, then it turns out that the square matrix $I - Q$ is always invertible. The matrix $N = (I - Q)^{-1}$ is called the *fundamental matrix* of the Markov chain: Let N_{ij} represent the ijth element of N, T_i the sum of the entries in row i of N, and B_{ij} the ijth entry of the matrix $B = N R$.

We can now state the answers to the four important questions about absorbing Markov chains:

1. Every absorbing Markov process eventually reaches some absorbing state.
2. The number N_{ij} is the average number of times the process is in the jth transient state if it starts in the ith transient state.
3. The number T_i is the average number of steps before the process enters an absorbing state if it starts in the ith transient state.
4. The number B_{ij} is the probability of eventually entering the jth absorbing state if the process starts in the ith transient state.

These results are easier to remember if you keep track of the sizes of the matrices. The matrix N is $(r-k)$ by $(r-k)$ and its rows and columns correspond to transient states. The matrix $B = N R$ has size $(r-k) \times k$; the rows correspond to transient states and the columns to absorbing states.

An Application

Consider the personnel policies of Lower Pine Cone College (Example 3) with the transition probabilities given in Section II.C:

$$a = .1, b = .01, c = .67, d = .22, e = .05, f = .45, g = 5.$$

The matrices R and Q are given by

$$R = \begin{pmatrix} .1 & .01 \\ .05 & .45 \end{pmatrix} \quad Q = \begin{pmatrix} .67 & .22 \\ .5 & 0 \end{pmatrix}$$

and we have

$$I - Q = \begin{pmatrix} 1 - .67 & 0 - .22 \\ 0 - .5 & 1 - 0 \end{pmatrix} = \begin{pmatrix} .33 & -.22 \\ -.5 & 1 \end{pmatrix}$$

so that

$$N = (I - Q)^{-1} = \begin{pmatrix} 4.54545 & 1 \\ 2.27273 & 1.5 \end{pmatrix}$$

and

$$B = NR = \begin{array}{c} \\ R \\ P \end{array} \begin{array}{c} \begin{array}{cc} F & T \end{array} \\ \begin{pmatrix} .504545 & .495455 \\ .302273 & .697727 \end{pmatrix} \end{array}$$

Here are some conclusions we may make about this absorbing Markov process:

(a) Every nontenured faculty member will eventually be promoted to a tenure rank or will be fired.

(b) A faculty member who has just been promoted but not given tenure (initial state P) will eventually be given tenure with probability .697727 (B_{22}) or be fired with probability .302273 (B_{21}). A professor should expect to wait, on average, 3.77273 years ($T_2 = N_{21} + N_{22}$) before the decision is made.

(c) A professor who was retained at her present rank by the committee this year (initial state R) faces a probability of .495455 of eventually gaining tenure and a probability of .504545 that she eventually will be fired. Her expected waiting time for a decision is 5.54545 years ($T_1 = N_{11} + N_{12}$).

You will find more extended discussions of absorbing Markov chains in the sports examples of Section IV.C below, in a mathematical model of learning in Chapter 13, and a model of recidivism in the criminal justice system in Chapter 17.

C. Sports Examples

Imagine two teams, A and B, playing a championship series of three games. The first team to win two games is declared the winner of the series. Let p represent the

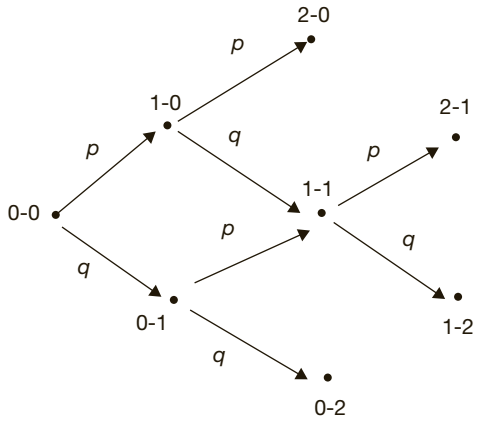


FIGURE 11.8 State diagram for a three-game series.

probability that team A will win any one particular game and let $q = 1 - p$. Let $m - n$ denote the state that the series stands at m wins for team A and n wins for team B. The initial state is then 0–0 while 2–0 denotes a clean sweep for A, and 1–1 means the series is even after 2 games. There are eight possible states: 0–0, 1–0, 0–1, 1–1, 2–0, 2–1, 1–2, 0–2. The absorbing states are 2–0, 2–1, 1–2, and 0–2 since the series ends as soon as one team has won two games. A state diagram for this Markov process appears in Fig. 11.8.

From the state diagram, a standard form for the transition matrix can be constructed:

$$\begin{array}{c} \\ 2-0 \\ 2-1 \\ 1-2 \\ 0-2 \\ 0-0 \\ 1-0 \\ 0-1 \\ 1-1 \end{array} \begin{array}{c} \begin{array}{cccccccc} 2-0 & 2-1 & 1-2 & 0-2 & 0-0 & 1-0 & 0-1 & 1-1 \end{array} \\ \left(\begin{array}{cccc|cccc} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & p & q & 0 \\ p & 0 & 0 & 0 & 0 & 0 & 0 & q \\ 0 & 0 & 0 & q & 0 & 0 & 0 & p \\ 0 & p & q & 0 & 0 & 0 & 0 & 0 \end{array}\right) \end{array}$$

The matrix $I - Q$ has the form

$$I - Q = \begin{array}{c} \\ 0–0 \\ 1–0 \\ 0–1 \\ 1–1 \end{array} \begin{array}{c} \begin{array}{cccc} 0–0 & 1–0 & 0–1 & 1–1 \end{array} \\ \begin{pmatrix} 1 & -p & -q & 0 \\ 0 & 1 & 0 & -q \\ 0 & 0 & 1 & -p \\ 0 & 0 & 0 & 1 \end{pmatrix} \end{array}$$

and the fundamental matrix $N = (1 - Q)^{-1}$ is

$$N = (I-Q)^{-1} = \begin{array}{c} \\ 0\text{–}0 \\ 1\text{–}0 \\ 0\text{–}1 \\ 1\text{–}1 \end{array} \begin{array}{c} \begin{array}{cccc} 0\text{–}0 & 1\text{–}0 & 0\text{–}1 & 1\text{–}1 \end{array} \\ \begin{pmatrix} 1 & p & q & 2pq \\ 0 & 1 & 0 & q \\ 0 & 0 & 1 & p \\ 0 & 0 & 0 & 1 \end{pmatrix} \end{array}$$

so that the matrix $B = NR$ is given by

$$B = NR = \begin{array}{c} \\ 0\text{–}0 \\ 1\text{–}0 \\ 0\text{–}1 \\ 1\text{–}1 \end{array} \begin{array}{c} \begin{array}{cccc} 2\text{–}0 & 2\text{–}1 & 1\text{–}2 & 0\text{–}2 \end{array} \\ \begin{pmatrix} p^2 & 2p^2q & 2pq^2 & q^2 \\ p & pq & q^2 & 0 \\ 0 & p^2 & pq & q \\ 0 & p & q & 0 \end{pmatrix} \end{array}$$

Since the series begins in state 0–0, examine the first rows of N and B. Team A wins the series if the absorbing state is 2–0 or 2–1. Thus, the probability that A wins the series is the sum of the first two entries of the first row of B:

$$\Pr(\text{Team } A \text{ wins series}) = p^2 + 2p^2q \tag{28}$$

The series lasts two games if the absorbing state is 2–0 or 0–2, and it lasts three games if the absorbing state is 2–1 or 1–2. From the first row of B, we obtain

$$\Pr(\text{Series lasts two games}) = p^2 + q^2 \tag{29}$$

$$\Pr(\text{Series lasts three games}) = 2p^2q + 2pq^2 = 2pq(p+q) = 2pq(1) = 2pq \tag{30}$$

The expected length of the series is the sum of entries in the first row of N:

$$1 + p + q + 2pq = 1 + p + (1-p) + 2p(1-p) = 2(1 + p - p^2) \tag{31}$$

Now the function $f(p) = 2(1 + p - p^2)$ achieves its maximum (2.5) at $p = .5$ and decreases monotonically to 2 as p increases to 1. Thus, we can determine p from the average length of a large number of series.

As an example, the U.S. Tennis Association's *Official Encyclopedia of Tennis* lists the results of 112 men's tennis tournaments in which the winner was determined by a three-set series. Of these matches, 67 were concluded after two sets and 45 lasted three sets, for an average length of 2.401 sets. From Eq. (31), this corresponds to a probability $p = .72486$ of a player winning a single given set.

With this value of p, Eqs. (29) and (30) predict that in 112 matches 67.22 would last two sets and 44.78 would last three sets.

Turn now to the situation of players A and B engaged in a five-set championship series. The first player to win three sets is the winner. As before, p denotes the probability that player A will win any one particular set and $q = 1 - p$.

Treating this series as a Markov process, there are 15 states:

Six absorbing states: 3–0, 3–1, 3–2, 2–3, 1–3, 0–3

Nine transient states: 0–0, 1–0, 0–1, 2–0, 1–1, 0–2, 2–1, 1–2, 2–2

The model of this competition is an absorbing Markov chain with matrices Q and R given by

$$Q = \begin{array}{c} \\ 0\text{–}0 \\ 1\text{–}0 \\ 0\text{–}1 \\ 2\text{–}0 \\ 1\text{–}1 \\ 0\text{–}2 \\ 2\text{–}1 \\ 1\text{–}2 \\ 2\text{–}2 \end{array} \begin{array}{c} \begin{array}{ccccccccc} 0\text{–}0 & 1\text{–}0 & 0\text{–}1 & 2\text{–}0 & 1\text{–}1 & 0\text{–}2 & 2\text{–}1 & 1\text{–}2 & 2\text{–}2 \end{array} \\ \begin{pmatrix} 0 & p & q & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & p & q & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & p & q & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & q & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & p & q & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & p & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & q \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & p \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \end{array}$$

and

$$R = \begin{array}{c} \\ 0\text{–}0 \\ 1\text{–}0 \\ 0\text{–}1 \\ 2\text{–}0 \\ 1\text{–}1 \\ 0\text{–}2 \\ 2\text{–}1 \\ 1\text{–}2 \\ 2\text{–}2 \end{array} \begin{array}{c} \begin{array}{cccccc} 3\text{–}0 & 3\text{–}1 & 3\text{–}2 & 2\text{–}3 & 1\text{–}3 & 0\text{–}3 \end{array} \\ \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ p & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & q \\ 0 & p & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & q & 0 \\ 0 & 0 & p & q & 0 & 0 \end{pmatrix} \end{array}$$

The fundamental matrix N is

$$\begin{array}{c} \\ 0\text{–}0 \\ 1\text{–}0 \\ 0\text{–}1 \\ 2\text{–}0 \\ 1\text{–}1 \\ 0\text{–}2 \\ 2\text{–}1 \\ 1\text{–}2 \\ 2\text{–}2 \end{array} \begin{array}{c} \begin{array}{ccccccccc} 0\text{–}0 & 1\text{–}0 & 0\text{–}1 & 2\text{–}0 & 1\text{–}1 & 0\text{–}2 & 2\text{–}1 & 1\text{–}2 & 2\text{–}2 \end{array} \\ \begin{pmatrix} 1 & p & q & p^2 & 2pq & q^2 & 3p^2q & 3pq^2 & 6p^2q^2 \\ 0 & 1 & 0 & p & q & 0 & 2pq & q^2 & 3pq^2 \\ 0 & 0 & 1 & 0 & p & q & p^2 & 2pq & 3p^2q \\ 0 & 0 & 0 & 1 & 0 & 0 & q & 0 & q^2 \\ 0 & 0 & 0 & 0 & 1 & 0 & p & q & 2pq \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & p & p^2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & q \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & p \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \end{array}$$

and the matrix $B = NR$ is

$$B = \begin{array}{c} \\ 0\text{–}0 \\ 1\text{–}0 \\ 0\text{–}1 \\ 2\text{–}0 \\ 1\text{–}1 \\ 0\text{–}2 \\ 2\text{–}1 \\ 1\text{–}2 \\ 2\text{–}0 \end{array} \begin{array}{c} \begin{array}{cccccc} 3\text{–}0 & 3\text{–}1 & 3\text{–}2 & 2\text{–}3 & 1\text{–}3 & 0\text{–}3 \end{array} \\ \begin{pmatrix} p^3 & 3p^3q & 6p^3q^2 & 6p^2q^3 & 3pq^3 & q^3 \\ p^2 & 2p^2q & 3p^2q^2 & 3pq^3 & q^3 & 0 \\ 0 & p^3 & 3p^3q & 3p^2q^2 & 2pq^2 & 0 \\ p & pq & pq^2 & q^3 & 0 & 0 \\ 0 & p^2 & 2p^2q & 2pq^2 & q^2 & 0 \\ 0 & 0 & p^3 & p^2q & pq & q \\ 0 & p & pq & q^2 & 0 & 0 \\ 0 & 0 & p^2 & pq & q & 0 \\ 0 & 0 & p & q & 0 & 0 \end{pmatrix} \end{array}$$

The sum of the entries in the first row of N gives the expected number of sets in the series:

$$1 + p + q + p^2 + 2pq + q^2 + 3p^2q + 3pq^2 + 6p^2q^2 \tag{32}$$

The sum $B_{11} + B_{12} + B_{13}$ gives the probability that player A wins the series:

$$\Pr(\text{A wins series}) = p^3 + 3p^3q + 6p^3q^2 \tag{33}$$

To determine the probability distribution for the length of the series, we have

$$\Pr(\text{Series ends in 3 sets}) = B_{11} + B_{16} = p^3 + q^3 \tag{34}$$

$$\Pr(\text{Series ends in 4 sets}) = B_{12} + B_{15} = 3p^3q + 3pq^3 \tag{35}$$

$$\Pr(\text{Series ends in 5 sets}) = B_{13} + B_{14} = 6p^3q^2 + 6p^2q^3 \tag{36}$$

To test this model of tennis competition, suppose that the probability $p = .72486$, based on our earlier observations, holds in general competition. Then the expected length of a best-of-five-sets series becomes, using Eq. (32),

$$\text{Predicted expected length} = 3.84 \text{ sets} \tag{37}$$

The national men's singles champion of the USTA is determined each year in a tournament that climaxes in a best-of-five-sets series between the two finalists. In the 133 championships decided between 1881 and 2013, there were

54 three-set matches
49 four-set matches
30 five-set matches

which gives

$$\text{Observed average length} = 3.82 \text{ sets} \tag{38}$$

The predicted and observed average lengths are remarkably close.

Furthermore, Eqs. (34)–(36) predict

53.4 three-set matches
47.8 four-set matches
31.7 five-set matches

in a group of 133 matches.

A standard statistical test (the chi-squared) shows that the observed and predicted distributions are significantly close together to lend weight to the assumption that tennis competition does follow the behavior of a Markov process.

Of course, the other rows of N and B also make predictions about the course of the tennis competition. The entry $B_{32} = p^3$, for example, predicts the probability of a player A winning the match if he loses the first set. We can compare this prediction to the observed value for the USTA championships as a further test of the model. The model predicts this would occur 8.23 times in a group of 89 matches. In actual fact, this has happened nine times in the USTA championships.

It is interesting to note that for a fixed probability p of winning any given set, the probability of winning a match increases with the number of required victories. Some representative figures are given in Table 11.2.

In their book on finite Markov chains, J. Kemeny and J. L. Snell (1960) investigate tennis competition from a slightly different point of view. As we have noted, a match is decided by the winner of a three-set or a five-set series. A player wins a set by being the first to win six or more games and have a lead of at least two games over his opponent. Thus, possible final scores in a set are 6–0, 6–1, . . . , 6–4, 7–5, 8–6, . . . , where the numbers represent games won. An individual game is won by the first player to amass 4 or more points, provided he leads by at least 2 points.

Kemeny and Snell compute the probability of winning a game, a set, and a match if a player has probability p of winning each *point*. They show, for example, that if $p = .51$,

Table 11.2 Probabilities of winning matches of different lengths for selected values of p, the probability of winning a single set.

Probability of winning a given set	.51	.6	.72846
Probability of winning 3-set match	.515	.648	.815
Probability of winning 5-set match	.519	.683	.868
Probability of winning 7-set match	.522	.710	.904

then the likelihood of winning the match is .635 while if $p = .6$, then the probability of winning the match rises to .9996. If there is a significant difference (.2 or more) in the abilities of the players, as measured by $p - (1 - p)$, then the better player is almost certain to win. Even if the difference is small ($p = .51$, $1 - p = .49$), the better player still wins more than 63% of the time.

D. Theorems

This section contains statements of and outlines of proofs for the major results about absorbing Markov chains already discussed and illustrated. Detailed, rigorous proofs may be found in Kemeny and Snell's book on Markov chains.

THEOREM 6 In an absorbing Markov chain, the probability that the process will eventually enter an absorbing state is 1.

Sketch of Proof of Theorem 6 Let S_i be a transient state of the Markov process. It is possible to reach at least one of the absorbing states in a finite number of steps, if the process begins in S_i. Let r_i denote the minimum number of steps necessary to reach some absorbing state from S_i. Let p_i denote the probability that the process does *not* reach any absorbing state in r_i steps if it starts in S_i. Then p_i is strictly less than 1.

Let r denote the largest of the numbers r_i and p the largest of the probabilities p_i where i ranges over the index numbers of the transient states. Then the probability of not reaching an absorbing state in r steps is less than p. Similarly the probability of not reaching an absorbing state in $2r$ steps is less than p^2. In general, the probability of not reaching an absorbing state in nr steps is less than p^n. Since $p < 1$, the probabilities p^n tend to 0 as n gets large. Thus, the probability of eventually reaching some absorbing state must tend to 1. ◇

THEOREM 7 Let P be the transition matrix of an absorbing Markov chain in standard form

$$P = \begin{array}{l} \\ k \text{ Absorbing States} \\ n-k \text{ Transient States} \end{array} \begin{array}{c} \begin{array}{cc} \text{Absorbing} & \text{Transient} \\ \text{States} & \text{States} \end{array} \\ \left(\begin{array}{c|c} I & 0 \\ \hline R & Q \end{array}\right) \end{array}$$

Then P^n has the form

$$P^n = \begin{pmatrix} I & 0 \\ R_n & Q^n \end{pmatrix}$$

where Q^n is the nth power of Q and R_n is a $(r - k) \times k$ matrix.

Proof of Theorem 7 Suppose the states have been numbered so that $S_1, S_2, \ldots, S_k$ are the absorbing states and $S_{k+1}, S_{k+2}, \ldots, S_r$ are the transient states. The matrix Q has the form

$$Q = \begin{pmatrix} q_{k+1,k+1} & q_{k+1,k+2} & \cdots & q_{k+1,r} \\ q_{k+2,k+1} & q_{k+2,k+2} & & q_{k+2,r} \\ \cdots & & & \\ q_{r,k+1} & q_{r,k+2} & & q_{r,r} \end{pmatrix}$$

where $q_{k+i,\,k+j}$ is the transition probability of moving from the ith transient state to the jth transient state in one step.

Consider the matrix P^2. The probability of moving from the ith transient state to the jth transient state in 2 steps is the $(k + i, k + j)$th entry of P^2. This number is the matrix product of the $(k + i)$-th row of P,

$$(r_{k+i,l}, r_{k+i,2}, \ldots, r_{k+i,k}, q_{k+i,k+1}, \ldots, q_{k+i,r})$$

and the $(k + j)$ th column of P,

$$\begin{pmatrix} 0 \\ 0 \\ \cdots \\ 0 \\ q_{k+1,k+j} \\ q_{k+2,k+j} \\ \cdots \\ q_{r,k+j} \end{pmatrix}$$

This product is equal to $\sum_{s=1}^{r-k} q_{k+i,k+s} q_{k+s,k+j}$ and that number is simply the product of the ith row and jth column of Q—that is, the ijth entry of Q^2. Thus, the lower right-hand corner of P^2 is Q^2. The corresponding result for P^n follows by induction on n. ◇

This theorem says that the probability of moving from the ith transient state to the jth transient state in n steps is the ijth entry of Q^n. From the proof of Theorem 6, however, we know that this probability tends to zero. Thus, we have

Corollary The powers of the matrix Q tend to 0—that is, $\lim_{n\to\infty} Q^n = 0$.

THEOREM 8 The matrix $I - Q$ has an inverse.

Sketch of Proof of Theorem 8 For $n = 1, 2, 3, \ldots$, let C_n be the matrix defined by

$$C_n = I + Q + Q^2 + \cdots + Q^{n-2} + Q^{n-1}$$

Then

$$QC_n = Q + Q^2 + \cdots + Q^{n-1} + Q^n$$

and

$$(I - Q)C_n = C_n - QC_n = I - Q^n. \tag{39}$$

Now let n increase on both sides of Eq. (39). Letting $N = \lim_{n\to\infty} C_n$ and using the fact that $\lim_{n\to\infty} Q^n = 0$, we have

$$(I - Q)N = I - 0 = I \tag{40}$$

so that N is the inverse of $I - Q$. ◇

Admittedly, the derivation of Eq. (40) from (39) is highly nonrigorous, but the reader is assured that everything can be fully justified with epsilons and deltas.

THEOREM 9 Let n_{ij} be the expected number of times that an absorbing Markov chain is in the jth transient state S_{k+j} if it starts in the ith transient state S_{k+i}. If N is the matrix whose entries are given by n_{ij}, then N is the inverse of $I - Q$.

Proof of Theorem 9 If $i = j$, then n_{ij} is at least one. Using the fact that the expected value of a sum is the sum of the expected values, we have the relation

$$n_{ij} = d_{ij} + P_{k+i,k+1}n_1j + P_{k+i,k+2}n_{2j} + \cdots + P_{k+i,r}n_{rj} \tag{41}$$

where $d_{ij} = 1$ if $i = j$ and 0 if $i \neq j$. The transition probabilities $p_{k+i,\,k+j}$ are obtained from the transition matrix. The relation (41) follows since we must consider all possible moves to other transient states at the first step. Note that the transition probabilities entering the equation are exactly the entries of Q. The matrix form of Eq. (41) then is

$$N = I + QN \tag{42}$$

which gives $IN = I + QN$ or $I = IN - QN = (I - Q)N$. Since $(I - Q)N = I$, N is the inverse of $I - Q$. ◇

Corollary The sum of the entries in any row of N is the expected number of times the process is in some transient state for a given starting transient state—that is, the expected number of steps before the process enters an absorbing state.

THEOREM 10 Let P be the transition matrix of an absorbing Markov chain in standard form and let $N = (I - Q)^{-1}$. Let b_{ij} be the probability that the process will enter the jth absorbing state if it starts in the ith transient state. Then b_{ij} is the ijth entry of the matrix $B = N\,R$.

Proof of Theorem 10 The process could enter the absorbing state at the first step with probability $P_{k+i,j}$ or it could first move to some other transient state and then eventually move into the jth absorbing state. Thus, we have the relation

$$b_{ij} = p_{k+i,j} + \sum_{s=1}^{r} p_{k+i,\,k+s} b_{sj} \tag{43}$$

where the summation runs over all the transient states. Note that the transition probability $p_{k+i,j}$ is an entry of R and the transition probabilities $p_{k+i,k+s}$ are entries of Q. Hence, the matrix form of Eq. (43) is

$$B = R + QB \tag{44}$$

so that

$$R = B - QB = (I - Q)B \tag{45}$$

Multiply each side of Eq. (45) by $N = (I - Q)^{-1}$ to obtain

$$NR = (I - Q)^{-1}(I - Q)B = B \tag{46}$$

◇

V. Historical and Biographical Notes

A. Markov

Novosti from Sovfoto

Andrei Andreevich Markov

Markov processes are named after the man who first studied them, the Russian mathematician Andrei Andreevich Markov. Markov was born in Ryazan, Russia, on June 14, 1856, and was the son of a member of the gentry who managed a large private estate.

The young Markov suffered from poor health and needed crutches until he was 10 years old. His talent for mathematics was spotted early and he received a gold medal for his undergraduate thesis at St. Petersburg University (1878), entitled "On the integration of differential equations by means of continued fractions." He completed his doctoral dissertation on continued fractions and the problem of moments 6 years later. In 1883, Markov married Maria Ivanovna Valvatyeva, whom he had known since childhood: she was the daughter of the proprietress of the estate managed by his father.

Markov combined an active research program with his teaching at St. Petersburg University for 25 years. He made important contributions to number theory, continued fractions, approximate quadrature formulas, function theory, integration in elementary functions, differential equations, and probability theory. He retired in 1905 to make room on the faculty for younger mathematicians, although he continued to present the course on probability. Markov's lectures and papers were noted for an irreproachable strictness of argument and a rather peremptory manner of stating opinions on the work of others. One biographer, Alexander Youschkevitch, described Markov as having "a mathematical cast of mind that takes nothing for granted" and reported that he was extremely exacting with his students and associates. It is said that during his lectures, Markov bothered little about the order of equations on the blackboard and even less about his personal appearance.

Markov was actively concerned with the politics of his time. He participated in the liberal movement in Russia at the beginning of the 20th century. He protested the Tsar's overruling of the election of Maxim Gorky to the St. Petersburg Academy of Sciences and repudiated his own membership in the electorate after the illegal dissolution of the Second State Duma (parliament) by the government in 1907. In 1913, when the government celebrated the 300th anniversary of rule by the Romanov family, Markov organized a countercelebration of the 200th anniversary of Bernoulli's discovery of the law of large numbers.

In September 1917, Markov asked to be sent to the interior of Russia. He spent the famine winter in the little country town of Zaraisk, teaching mathematics without pay. He died in St. Petersburg on May 20, 1922.

Although he worked in a number of different areas of mathematics, Markov's contributions to probability theory produced the greatest effect on the development of science. The work on the law of large numbers and the central limit theorem by Chebyshev (Markov's teacher), Lyapunov, and Markov created the basis for the modernization of probability theory.

Markov initiated the study of stochastic processes that would later bear his name in a 1906 paper "Rasprostranenie zakona bolshikh chisel na velichiny, t zavi syaschchie drug ot druga" ("The Extension of the Law of Large Numbers on Mutually Dependent Variables"). Markov arrived at his chains by starting from the internal needs of probability theory and not from applications to the physical or social sciences. He did study the application of his theory to the distribution of vowels and consonants in Pushkin's *Eugene Onegin*; this work is often cited as the first modern paper on mathematical linguistics.

B. Further Applications

The English biologist-mathematician Sir Francis Galton (1822–1911) became interested in the survival and extinction of family names. The mathematical model he formulated in 1889 was one that involved the fundamental assumptions of Markov processes.

Paul and Tatyana Ehrenfast investigated a Markov chain model for diffusion in a 1907 paper about the same time that Einstein and Smoluchowski were using Markov processes to study Brownian motion.

Since these early studies, there have been many applications of Markov processes to the modeling of phenomena in the physical, life, and social sciences. Physicists have employed them to the theory of cascade processes, radioactive transformation, nuclear fission detectors, and the theory of tracks in nuclear research emulsions. Astronomers have studied fluctuations in the brightness of the Milky Way and the spatial distribution of galaxies using Markov chains. Chemists use stochastic models to understand chemical reaction kinetics and the statistical theory of polymer chains.

Mathematically inclined biologists have employed Markov chains and more general stochastic processes to learn more about population growth, structure of biological populations, taxis and kinesis, embryogenesis, evolution, molecular genetics, pharmacology, tumor growth, and epidemics.

Some of the areas of investigation in the social sciences that have been pursued through the use of Markov chains include voting behavior, geographical mobility within a country, the spread of ghettos in urban areas, growth and decline of towns, competition in the brewing industry, the size of economic firms, the spread of the use of intrauterine devices in Taiwan, prediction of enrollments in colleges and universities, the epidemiology of mental diseases, changes in personal attitudes, and the deliberations of a trial jury.

EXERCISES

I. Markov Chains

1. Two competing companies, Pollution Products and Environmental Hazards, simultaneously introduce new enzyme laundry detergents. Market tests indicate that during a year, Pollution keeps 60% of its customers and loses 40% of its customers to Environmental. On the other hand, Environmental keeps half of its customers and loses the other half to Pollution. Set up this process as a Markov chain. Determine the transition matrix and sketch a state diagram.

2. Abigail spends her entire weekly allowance on either candy or toys. If she buys candy 1 week, she is 60% sure to buy toys the next week. The probability that she buys toys in two successive weeks is 1/5. Set up this process as a Markov chain. Determine the transition matrix and sketch a state diagram.

3. A political scientist in Canada discovered that of the children of Conservatives, 80% vote Conservative and the rest vote Labor; of the sons and daughters of Labor supporters, 60% vote Labor, 20% vote Conservative, and 20% vote for the New Democratic Party (NDP); and of the offspring of NDP followers, 75% vote NDP, 15% vote Labor, and 10% vote Conservative.

(a) What is the probability that the grandchild of a Conservative will vote for the NDP?

(b) Set up this process as a Markov chain, with steps corresponding to successive generations. Determine the transition matrix and sketch the state diagram.

4. A secret CIA report gives the following analysis of the arms race between India and Pakistan: There are four possible states: War, Total Disarmament, Escalating Arms Race, and De-escalating Arms Race. It is not possible to change the situation if War or Total

Disarmament is occurring this year. If there is an Escalating Arms Race this year, the probability of continued escalation next year is .6, of de-escalation next year is .2, and of War next year is .2. If there is a De-escalating Arms Race this year, the probability of continued de-escalation next year is .7, of escalation next year is .1, and of total disarmament next year is .2.

Set up this process as a Markov chain. Determine the transition matrix and sketch the state diagram.

5. The National League and American League used to alternate as hosts for the opening game of each year's World Series. Show that this process can be set up as a Markov process with transition matrix $P = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$.

6. A particle moves along a line from an initial position 2 feet to the right of the origin. Each minute it moves one foot to the right with probability 1/2 or 1 foot to the left. There are barriers at the origin and 4 feet to the right of the origin; if the particle hits a barrier, it remains there. Show that this process can be set up as a Markov process with five states. Determine the transition matrix and draw the state diagram.

7. The random walk of Exercise 6 is modified so that if the particle reaches the barrier at the origin, it must move one foot to the right in the next minute, while if it hits the other barrier, it must move one foot to the left in the next minute. Determine the transition matrix for the associated Markov chain.

8. Find the probability that a woman whose birth weight was average has a granddaughter with an average birth weight.

9. Sketch the state diagram for the Lower Pine Cone College example.

10. Can accurate weather predictions be made from a Markov model of climate that uses only today's weather to forecast tomorrow's?

II. Matrix Operations and Markov Chains

11. Using probabilistic considerations only, show that the square of a stochastic matrix is also a stochastic matrix.

12. A stochastic matrix is *doubly stochastic* if the sum of the entries in each column is 1. If A and B are doubly stochastic square matrices, is A^2 doubly stochastic? Is AB?

13. Write out inductive proofs for Theorems 2 and 3.

14. Use matrix multiplication to solve Exercise 8.

15. Abigail bought a toy with her allowance this week (see Exercise 2). Find the probability that she will buy a toy 4 weeks from now.

16. Find the distribution of birth weights (Example 1) after one generation if the initial probability distribution is (.4, .3, .3).

17. Suppose the distribution of birth weights of a generation of daughters is $p^{(1)} = 7(.31, .45, .24)$. Can you find the distribution of birth weights of the mothers?

18. Use the state diagram in Fig. 11.3 to find the probability that the process reaches state S_3 in two steps if it starts in state S_1.

19. Let P be the transition matrix of Eq. (4). Compute P^2. What can you say about the first column of P^3? P^n?

20. Consider the transition matrix for the Lower Pine College model. Can you determine whether it is more likely that a newly hired professor will eventually be given tenure or be fired?

III. Regular Markov Chains

21. The matrices determined in Exercises 1–7 are all transition matrices for Markov processes. Which ones are regular?

22. Find, if possible, a fixed-point vector for each of the transition matrices of Exercises 1–7.

23. Does a matrix of the form $A = \begin{pmatrix} a & b \\ -b & d \end{pmatrix}$ have a fixed-point vector?

24. How often, in the long run, will a stereo be the special sale item at the Audio-Visual Den (Example 2)?

25. Analyze the long range prospects for the competition model described in Exercise 1.

26. How often, on the average, does Abigail spend her allowance on candy (Exercise 2)?

27. What are the long-term predictions for the male vote in Canada according to the data of Exercise 3?

28. Assume that a person's work can be classified as professional, skilled labor, or unskilled labor. Assume that of the children of professionals, 80% are professional, 10% are skilled laborers, and 10% are unskilled laborers. In the case of children of skilled laborers, 60% are skilled laborers, 20% are professionals, and 20% are unskilled laborers. Finally, in the case of unskilled laborers, 50% of the children

are unskilled laborers, and 25% each are in the other two categories. Assume that every person has a child, and form a Markov chain by following a given family through several generations. In commenting on the society described, the famed sociologist Harry Perlstadt has written, "No matter what the initial distribution of the labor force is, in the long run the majority of the workers will be professionals." Is he correct? Why?

29. Suppose P is the transition matrix of a regular Markov chain and let W be the matrix given by Theorem 4. Prove that the matrix $I-(P-W)$ is invertible. Compute this matrix and its inverse if P is the matrix of Example 5.

30. Let $N=\Big(I-(P-W)\Big)^{-1}$ be the inverse of the matrix of Exercise 29. Here N is called the *fundamental matrix* of a regular Markov process. Show that each of the following statements is true (a) if P is the matrix of Example 5 and (b) P is the transition matrix of any regular Markov process:

(i) $NP=PN$.

(ii) $\mathbf{w}N=\mathbf{w}$, where w is the fixed-point stochastic vector of P.

(iii) $I-N=W-PN$.

(iv) N is a stochastic matrix.

31. Let u be a given stochastic vector. Is it always possible to find a regular transition matrix P such that u is a fixed-point vector of P?

IV. Absorbing Markov Chains

32. Which of the transition matrices of Exercises 1–7 represent absorbing Markov processes?

33. What is the likelihood that the arms race described in Exercise 4 will end in a war? If there is an escalating arms race this year, what is the expected number of years before the arms race resolves itself into war or disarmament?

34. What predictions can you make about the random walk models of Exercises 6 and 7 using the theorems about regular and absorbing Markov processes?

35. Let N be the fundamental matrix of an absorbing Markov chain. Show that N is invertible, and prove that $NQ=N-I$.

36. Let S_i and S_j be two transient states in a Markov process. Show that there are positive numbers b and c, with c less than 1, so that $p_{ij}^{(n)}\le bc^n$. [*Hint*: Examine the proof of Theorem 6.]

37. Let P be the transition matrix of an absorbing Markov chain with r states. Let B^* be the $r\times r$ matrix whose ijth entry is the probability of being absorbed in state S_j if the process starts in state S_i. Prove that $PB^*=B^*$.

38. Supply a rigorous proof for Theorem 8.

39. A trio of 19th century Russian noblemen fight a three-way duel. The three men are of different abilities at pistol shooting. They have respective probabilities of 1/2, 1/3, and 1/6 of hitting and killing the target at which they aim. In each round of the duel, the men shoot simultaneously and each one aims at the best marksman not yet killed. Treat this duel as a Markov chain by taking as the states the men who survive any one round. Find N and B and interpret the results.

40. *Gambler's Ruin*. Annie has \$3 and Rachel has \$2. They flip a fair coin. If it is a head, Annie pays Rachel \$1. Otherwise, Rachel pays Annie \$1. How long will it take for one of the players to go broke or win all the money?

V. Further Applications

For the remaining problems, see the last section of Chapter 10.

41. Each individual belongs to one of three possible genotypes: *AA* (dominant), *AB* (hybrid), or *BB* (recessive). In a laboratory experiment, an individual of unknown genotype is mated with a hybrid. Show that the probabilities for the genotypes *AA*, *AB*, and *BB* of the offspring are given by vectors:

(a) (.5, .5, 0) if the unknown parent is dominant

(b) (.25, .5, .25) if the unknown parent is hybrid

(c) (0, .5, .5) if the unknown parent is recessive

42. Suppose the experiment of Exercise 41 is repeated a large number of generations—that is, in each generation an offspring is chosen at random and mated with a hybrid. Set up this process as a Markov chain, show that it is regular, and find the unique fixed-point stochastic vector. Interpret the result.

43. Consider a large population in which mating is completely random, half the offspring are female, and the proportion of genotypes is the same for both males and females. Let $p^{(0)}$ be the probability vector for the genotypes of an initial generation of parents and let $p^{(k)}$ be the probability vector of genotypes of the kth generation of offspring.

(a) Find the transition matrix P of this Markov process so that $p^{(k)}=p^{(0)}P^k$.

(b) Show that $p^{(1)}$ is a fixed-point vector for P.

The conclusion of (b) is that the distribution of the genotypes is stable after only one generation. This result is called the *Hardy-Weinberg equilibrant principle* and was discovered independently in 1908 by G. H. Hardy and W. Weinberg.

44. Show that the Hardy-Weinberg principle may not be valid if parents of one genotype have, on the average, more offspring than parents of another genotype. How can the principle be modified in such a case?

SUGGESTED PROJECTS

1. Analyze World Series competition in the spirit of the tennis examples as an absorbing Markov chain. Let p be the probability of winning any particular game. Determine the transition matrix P and associated matrices Q, N, and B. Show that for $.5 \leq p \leq 1$, the expected length of a World Series is a monotonically decreasing function p. Thus, p can be determined from the observed average length of World Series competition. Does this value of p predict closely the number of 4-, 5-, 6-, and 7-game series that have occurred? Is there some way of estimating p without relying on World Series information? Here are some possibilities:

(a) Let A be the average number of runs scored in the season by the American League pennant winner, and let L be the similar number for the National League counterpart. Let p, the probability that the American League champion wins a given game, be $A/(A + L)$.

(b) Instead of using runs scored, use the difference between runs scored and runs allowed.

(c) Instead of using runs scored, use the number of games won.

2. An *ergodic* Markov process is one in which it is possible to go from any given state to any other one in a finite number of steps, but the number may depend on which states are chosen. For example, certain states may only be reached in an odd number of steps, while others require an even number. Show that every regular Markov process is an ergodic one. Find some examples of ergodic chains that are not regular. Many of the important theorems about regular chains are also true for ergodic ones, although of necessity, the proofs are different. Which theorems are these? What real-world processes can be modeled with ergodic Markov chains?

3. Markov models have been used in many studies of learning theory and social conformity in which the states correspond to certain "states of mind," which may not be directly observable: the subject may be limited to a certain number of observable responses, for example, but the same response can occur if the subject is in any of several different states. Can the transition matrix be reconstructed from the observed behavior? Assuming that the initial state can be determined, note that the entries of the matrix B can be observed. Is this information enough to find Q and R?

4. It may happen that a stochastic process operates in such a fashion that the probability of being in a particular state at any step depends on the states occupied in the *two* immediately preceding steps. Strictly speaking, this is not a Markov process. Show that it can be made into a Markov process by doubling the number of states. One area where this idea can be applied is to the study of the outcomes of political elections where the winners come from one of several parties. Would a model assuming that the results of an election depend on the two preceding elections necessarily be a more accurate one than a model taking into account only the most previous election? Use such a model to study the results of Congressional elections in a single district during the 20th century.

5. Consider a Markov chain with transition matrix P. Suppose that before making a transition from state S_i to state S_j the process spends a time t_{ij} in state S_i. These *holding times*, t_{ij}, may be given by a probability distribution. Such a process is called a *semi-Markov process*. It has been used as a model to study the movement of coronary patients within different care units of a hospital (the absorbing states are death and discharge from hospital). Let t_i be the time the process spends in state S_i for each transition into that state. Find the expected value of t, the expected total amount of time the process will spend in state S_i if it has just arrived in state Si, and the expected amount of time the process will spend in transient states.

6. Investigate how the following two questions could be answered for a regular or an absorbing Markov process: If S_i, S_j and S_k are any three states,

(a) What is the expected number of steps for the process to move from S_i to S_j for the first time?

(b) If the process starts in S_i, what is the probability that it will reach S_j before S_k?

You can find a listing of references and suggestions for additional reading on the book's website, www.wiley.com/college/olinick

Two Models of Cultural Stability

CHAPTER 12

> Anthropology needs mathematics, not because mathematics is glamorous these days, but because mathematics can help anthropologists . . . solve the kinds of problems anthropologists want to solve.
>
> —Paul Kay

I. Introduction

Communities of people cannot long survive unless the basic needs of the inhabitants are met. In a rudimentary "society" each individual might take care only of his own requirements. He would find food, gather and prepare it, build his own shelter, and provide his own entertainment, medical care, and transportation. In most societies, however, people are dependent on one another for various goods and services. There is a division of labor among the residents. One person or group of persons specializes in constructing houses while another harvests the crops. Certain members debate and modify the laws, while others ensure that violators are apprehended and punished.

Furthermore, the obligations and the privileges of a single member of the society are different at different stages of the person's life. The social and economic contributions to the community of a 7-year-old, for example, vary from that of a 47-year-old. These in turn are not the same as those of a person of age 77.

Other factors besides age are often important in determining what is expected of an individual or what she is allowed to do. The person's gender, race, sexual orientation, religion, and perhaps even height and weight can control what occupations she will pursue and the extent of her power or influence in the society. The continued cultural viability of a community may depend quite crucially on the factors that are used to structure the division of privilege and responsibility among the members.

II. The Gadaa System

In this chapter, we will examine a system, the *Gadaa* system, for the division of labor that has been used by some of the peoples in East Africa. The Oromo, meaning "free men," constitute one of the largest racial groups in Ethiopia and a small minority in Kenya. Historically, the Oromo were called Galla, a name now considered pejorative. The Bilisummaa Oromiyaa

(Oromiyaa Liberation Council) highlights the centrality of the Gadaa system for their society [2007]:

> *We Oromos have taken onto our own shoulders the crucially important task of beginning, encouraging, and carrying out the reconstruction of Oromo history—i.e., uncovering data, recording and analyzing the events and finding clues buried in the past and present about the very structure of Oromo society. There is no question that central to any study of the Oromo is the GADA system. GADA is recognized by all Oromos as a key to the unique heritage of Oromo political, social, and cultural life. Whereas most of us do know about the existence of the GADA system from our elders, its specific operations are unclear to most of us. For this reason Oromo intellectuals have decided to spend significant time and energy on the study of GADA. Our study so far has led us to suggest that a beacon and even a blueprint for democracy in Oromia may be found in the kind of society that Oromos maintained in the past and have preserved in various forms into the present. The GADA system is a key since it has been the predominant organizational form in Oromo society.*

We shall consider only a simplified version of the actual system, so that we may focus on some important questions. In the Gadaa system, the critical functions of the tribe are structured through five age grades, called the Dabella, Folie, Kondala, Luba, and Yuba.

Each male in the society moves through the age-grade system, spending a period of 8 consecutive years in each grade. Since there are five grades, it takes an individual 40 years to pass through the system. The key feature of this age-grade system is that a man enters the lowest grade at the moment his father retires from the highest grade. In other words, a son enters the system exactly 40 years after his father enters.

To illustrate this scheme, suppose your father enters the lowest grade when he is 13 years old and that you are born when he is 30 years old. Then your father retires from the system when he is 53 $(40 + 13)$ years old. At that time, you will enter the system. Your age will be $53 - 30 = 23$.

To continue this example, suppose that you have two sons, one who is born when you are 35 and the other when you are 45. You will leave the system at age 63. Your two sons will enter the system at the same time, although their ages will be different. The elder will be 28 years old and the younger will be 18 years old. They will move through the system together, entering the successive grades at the same time, and retiring in the same year.

The calendar years of entrance into the Gadaa system of all the male descendants of a man is then determined once we know the year the man himself entered. If a man enters the lowest grade in the year 2015, all his sons enter in the year 2055, even if the man dies before the date of his retirement. It is also possible that a son may enter the system before he is born! To see how this may happen, suppose that the man who entered the system in 2015 was very young. Then it is quite conceivable that he has a son who is born in 2065. In such a case, it is assumed that the son entered the lowest (Dabella) grade in 2055. At his birth then, the son is considered to be a member of the Folie grade and will advance to the Konda grade in 2061. This son would retire from the grade system in 2095 at the age of 30.

This age-grade system, as we have described it so far, poses no essential problems for the Oromo society. What makes the system interesting to study is that the roles of a male in the tribe depend entirely on which of the five grades he is occupying and not on his age,

wisdom, or strength. Two members of the Luba grade, for example, have the same rights and responsibilities, even if one is 7 years old and the other 47.

What are the particular roles assigned to the males in each grade? The anthropologist George Peter Murdock [1959] gives a concise description:

> *During the first grade . . . males are forbidden to have sex relations and they wander about begging food, which is always termed "milk" from married women. This is strongly suggestive of the behavior of infants. During the second grade they become initiated into sexual life but without forming stable relationships, and they engage in masked processions and behave generally in an irresponsible manner suggestive of adolescence. In the third grade they serve as warriors and are permitted to marry. Military valor is encouraged in some tribes . . . by requiring the taking of the genitals of a slain enemy as a trophy to qualify for full participation in the activities of the next, or ruling, grade. When an age-set enters the fourth, or Luba, grade, its members take over all important administrative, judicial, and priestly offices in the tribe and run its affairs for eight years. . . . The chief of the age-set, elected when it occupied the second grade, now becomes the high chief of the tribe. Another man becomes speaker of the general assembly. Others assume various administrative and judicial offices—chief priest, finance minister, and so on. During the last, or Yuba, grade, these men relinquish their posts and become "guardians," serving the new officials in a purely advisory capacity.*

Murdock's description indicates that the system may have been based, in its origin, on the maturity levels and abilities that corresponded with chronological age—that is, when the age-grade system began, the lower grade was made up entirely of children, while the highest grade was composed of the tribe's elders. As we have seen from our examples, however, in succeeding generations, the relationship between a man's age and the grade he occupies may be very complex.

Photograph by Herbert S. Lewis. Used with permission of Professor Lewis

FIGURE 12.1 Oromo ceremony of laying down of spears to settle a murder case. Herbert S. Lewis, photographer; used by permission of University of Wisconsin at Madison.

Since the rules of the age-grade system permit a young man to occupy a high grade, while an older man may be restricted to the activities allowed to members of a lower grade, tensions can easily arise in the tribe. Another anthropologist, Hans Hoffmann [1965], studied this system with the use of mathematical models. He raised the fundamental problem:

> *It is evident that the stability of Galla communities is threatened by the arbitrary interval of 40 years that is interposed between generations. Since this interval is often greater than the actual chronological difference between generations, the ages of some of the people in the grades may become progressively greater. This can result in humiliation and incongruity. An old man, entering the first grade, would be required to abstain from sexual activity and to wander around with its youthful members begging food. Further, if he should die before attaining the higher grades, important governmental offices may go unfilled.*

The fact that it is possible for a man to "enter" the age-grade system before his actual birth leads to a similar kind of difficulty. He may reach the middle grades of the system at too early a chronological age. He may not be equipped to fulfill the military or ruling functions with any competence. By the time he has the physical strength, talents, and experience to occupy these roles with distinction, he has graduated to the highest grade, where his services are no longer available to the tribe. Thus, the society may be seriously weakened, because it does not have access to the skills of its members at the time it needs them.

"It is curious fact," writes A. H. J. Prins [1953], "peculiar to the Galla institution, that the physical age of those who occupy simultaneously one and the same grade varies so widely, even from young children to fairly old men. . . . Viewed from the institutional angle, what it comes to is that most members of any grade fail to accomplish what is socially expected of them. The grade is supposed to exist because of the expected execution of a delegated task which regards more or less the real ages of the participants, but owing to factual circumstances widely differing from those reflected in the implicit charter, the grades, especially the lower ones, seem to have become an institutional (or even 'functional') failure. This failure has to be attributed to the composition of the personnel."

If there is too high a proportion of males in the society who are "out of phase" with the roles of the age-grade system, it will be difficult to maintain both a strong community and the age-grade system.

Since every community must place a high premium on its own survival, we may well ask if the age-grade system can continue unchanged over a period of many generations. Does the system possess stability as a component of the culture? Must it change to relieve the tensions we have described? Or will the differences between the 40-year intervals and the gaps between successive generations somehow "smooth out" over the years so that these tensions are essentially absent?

Hoffmann developed two models to study these questions. The first [1965] is a relatively simple deterministic model, while the second [1971] is a more sophisticated probabilistic one that makes use of Markov chains.

III. A Deterministic Model

To formulate a mathematical model, we must make some careful definitions and assumptions about the phenomena we hope to study. To investigate whether the age-grade system

possesses stability, Hoffmann [1965] first had to make precise the idea of a stable system. He proposed the following:

DEFINITION A *stable* system is one that tends to maintain a realistic relationship between age and role behavior.

For his deterministic model, Hoffmann investigated an axiom about sufficient conditions for a system to be stable.

AXIOM 1 A realistic relationship between age and role behavior can be maintained if, between any arbitrary number of generations, the ages at which an ancestor and his distant offspring entered the first grade are equal.

This axiom provides the means for translating our verbal discussions about stability into mathematics. Note that the condition for stability is an equality between numbers. We can make this more transparent by introducing some notation.

For the *i*th generation, we let A_i denote the age at which a man enters the first grade. Thus, A_1 gives the age of the first man of interest when he enters the lowest grade, A_2 the age of his son, A_3 the age of his grandson, and so on. For simplicity, we will assume that each man has exactly one son.

By P_i, we will denote the age of the man in the *i*th generation when his son is born.

In terms of this notation, we have two ways of writing the age of the man in the first generation at his retirement from the age-grade system. On the one hand, since he enters the system at age A_1 and remains in it for 40 years, he retires at age $A_1 + 40$. On the other hand, since his son enters at the time the father retires, the father's age at retirement is also given by $P_1 + A_2$. Thus, we have the basic relationship,

$$A_1 + 40 = P_1 + A_2 \tag{1}$$

which we may rewrite as

$$A_2 = A_1 + 40 - P_1 \tag{2}$$

If we require that a man and his son enter the age-grade system at the same age, then we are insisting that $A_1 = A_2$. Substituting this equality into Eq. (2) yields

$$40 - P_1 = 0$$

or

$$P_1 = 40.$$

The basic relationship stated in Eq. (2) holds for every pair of father and son; thus, we have

$$A_{i+1} = A_i + 40 - P_i \tag{3}$$

In particular, this gives us

$$\begin{aligned} A_3 &= A_2 + 40 - P_2 \\ &= (A_1 + 40 - P_1) + 40 - P_2 \\ &= A_1 + 2(40) - (P_1 + P_2) \end{aligned} \tag{4}$$

and

$$\begin{aligned} A_4 &= A_3 + 40 - P_3 \\ &= A_1 + 2(40) - (P_1 + P_2) + 40 - P_3 \\ &= A_1 + 3(40) - (P_1 + P_2 + P_3) \end{aligned} \tag{5}$$

A simple induction argument shows that

$$A_{n+1} = A_1 + n(40) - (P_1 + P_2 + \cdots + P_n) \tag{6}$$

We have seen that a man and his son will enter the age-grade system at the same age exactly if the son is born when his father is 40 years old. From Eq. (4), we may conclude that a man and his grandson will enter the age-grade system at the same age—that is, $A_3 = A_1$ exactly if

$$P_1 + P_2 = 80 \tag{7}$$

or

$$\frac{P_1 + P_2}{2} = 40 \tag{8}$$

This last equation asserts that the average age of parenthood of the first two generations must be 40 if the man and his grandson are to enter the system at the same age.

Now, Axiom 1 asserts that stability is maintained if the ages of entry of a man and his distant descendant are the same. If n denotes a large, arbitrary number of generations, then this condition is expressed by the equality

$$A_{n+1} = A_1 \tag{9}$$

Substituting this equality into Eq. (6) gives

$$A_1 = A_1 + 40n - (P_1 + P_2 + \cdots + P_n) \tag{10}$$

or

$$40n = P_1 + P_2 + \cdots + P_n \tag{11}$$

so that

$$\frac{P_1 + P_2 + \cdots + P_n}{n} = 40 \tag{12}$$

Now the number on the left-hand side of Eq. (12) is simply the average of the numbers $P_1, P_2, \ldots, P_n$. We may then conclude from this deterministic model that the age-grade system of the Gadaa will be stable if, over a large number of generations, the average age at which a man becomes a father is 40.

This deterministic model has the advantage that the predicted condition for stability—average age of 40 for parenthood—can be readily checked by examining accurate census data for the tribe.

This model has a number of important limitations, however. It deals only with one-dimensional father-son links and ignores the branching of descent lines representing siblings. In other words, the model assumes that a man has only one son when, in fact, many men have several sons. Of course, some men have no sons, and this shows another weakness of the model. The model transforms every given family into points of future time when it may, in fact, no longer exist.

Other aspects of this model and possible refinements and improvements of it will be presented in the exercises. In the next section, we will examine Hoffmann's probabilistic model for the question of cultural stability of the age-grade system.

IV. A Probabilistic Model

The stability of the Oromo age-grade system is threatened by the possibilities of disparities between chronological ages and assigned cultural roles. To promote stability, it is desirable that the lower grades consist largely of adolescents. What is crucial is not the absolute number of members of different ages in a particular grade, but the relative numbers. If most candidates for initiation into the lower grades are youthful, then there will be little tension in the system and we may expect it to continue to function largely unchanged for a number of generations. We can predict the level of tension that is likely to arise in the future if we know the ages of the males at the time they enter the age-grade system.

For computational simplicity, we will consider three age categories, or states, for the age at the time of initiation into the lowest grade:

S_1: ages 13−19

S_2: ages 20−29

S_3: ages 30 or over

The vector (x_1, x_2, x_3) will represent the proportion of males in each state. For example, if there are 100 men about to be initiated into the age-grade system with 25 in S_1, 55 in S_2, and 20 in S_3, we will represent this by the vector

$$(25/100, 55/100, 20/100) = (.25, .55, .20) \tag{13}$$

If the Gadaa system is to survive, the set of males about to enter the grade system should consist largely of younger men. The state S_1 should contain a relatively large proportion of the set while S_3 should include a relatively smaller fraction. As new generations enter the system, there should not be a significant drift from S_1 to S_3.

To determine the shifts from one state S_i to another S_j in successive generations, we determine for each male about to enter the system, the age of his father when the father entered the system.

Using Hoffmann's example, suppose that we examine a set of 240 males and record for each his state and his father's state at the time of initiation. The data are conveniently displayed in a matrix in which the rows correspond to the father's state and the columns to the son's state:

$$\begin{array}{rc} & \text{Son's state at time of initiation} \\ & \begin{array}{ccc} S_1 & S_2 & S_3 \end{array} \\ \begin{array}{r}\text{Father's} \\ \text{state at time} \\ \text{of initiation}\end{array} & \begin{array}{c} S_1 \\ S_2 \\ S_3 \end{array} \begin{pmatrix} 10 & 25 & 30 \\ 55 & 60 & 35 \\ 5 & 15 & 5 \end{pmatrix} \end{array}$$

We see from this matrix that the largest group (60) were in their twenties when they were initiated and so were their fathers. There were only five males who were initiated after the age of 30, but whose sons were initiated in their teens.

As noted above, our concern is not so much with absolute numbers, but with proportions. If we examine the 65 fathers who were initiated into the age-grade system as teenagers, we see that 10/65 of them had sons who were initiated as teenagers, 25/65 had sons initiated in their twenties, and 30/65 had sons initiated after the age of 30. We compute similar fractions for the fathers in states S_2 and S_3 and obtain the matrix

$$\begin{array}{c} \\ S_1 \\ S_2 \\ S_3 \end{array} \begin{array}{c} \begin{array}{ccc} S_1 & S_2 & S_3 \end{array} \\ \begin{pmatrix} 10/65 & 25/65 & 30/65 \\ 55/150 & 60/150 & 35/50 \\ 5/25 & 15/25 & 5/25 \end{pmatrix} \end{array} = \begin{array}{c} \\ S_1 \\ S_2 \\ S_3 \end{array} \begin{array}{c} \begin{array}{ccc} S_1 & S_2 & S_3 \end{array} \\ \begin{pmatrix} 2/13 & 5/13 & 6/13 \\ 11/30 & 12/30 & 7/30 \\ 1/5 & 3/5 & 1/5 \end{pmatrix} \end{array}$$

or, in decimal notation,

$$p = \begin{array}{c} \\ S_1 \\ S_2 \\ S_3 \end{array} \begin{array}{c} \begin{array}{ccc} S_1 & S_2 & S_3 \end{array} \\ \begin{pmatrix} .154 & .384 & .462 \\ .367 & .4 & .233 \\ .2 & .6 & .2 \end{pmatrix} \end{array}$$

Now it is possible to regard the entries in the matrix as probabilities. Thus, the probability that a father in state S_2 has a son in state S_3 is given as .233. Hoffmann considers this matrix a transition matrix from one generation to the next and notices that if we assume that this matrix remains constant, then we can study the age-grade system using the tools of Markov chain analysis.

For example, if our initial distribution of states is given by the vector

$$\mathbf{p}^{(0)} = (.25, .55, .20)$$

then the distribution of states after one generation is

$$\mathbf{p}^{(1)} = \mathbf{p}^{(0)} P = (.28, .44, .28)$$

After a single generation, there will be a slightly higher proportion of sons in S_3 than there were sons a generation ago. If this trend continues, the stability of the age-grade system is threatened.

Let's see what happens to the distribution after two generations. It will be given by the vector $\mathbf{p}^{(2)}$ where

$$\mathbf{p}^{(2)} = \mathbf{p}^{(1)} P = (\mathbf{p}^{(0)} P)P = \mathbf{p}^{(0)} P^2 = (.27, .46, .27)$$

It is easy to see that the drift from S_1 to S_3 evidenced after one generation has not continued.

In a similar fashion, we can compute the distribution of states after three generations, four generations, and so on. It is more interesting at this point, however, to determine the long-range behavior of the distribution vector. All the entries of the transition matrix P are positive, so we are dealing with a regular Markov process. The long-term distribution of states is then given by the unique fixed-point stochastic vector $\mathbf{w}$ of P. This is the vector $\mathbf{w} = (w_1, w_2, w_3)$ with the properties that $\mathbf{w} = \mathbf{w}P$ and $w_1 + w_2 + w_3 = 1$. Using the methods of Chapter 11 and Appendix II, we find that the components of $\mathbf{w}$ are

$$w_1 = \frac{663}{2518} = .263$$

$$w_2 = \frac{1140}{2518} = .453$$

$$w_3 = \frac{715}{2518} = .284$$

If our process is a Markov chain, then the proportion of males in the three states will tend toward $S_1 = .263$, $S_2 = .453$, $S_3 = .284$. Hoffmann notes that these values are not radically different from the initial vector $(.25, .55, .20)$ so that the system may be considered stable.

V. Criticisms of the Models

In what sense is Hoffmann correct in claiming that the mathematical model predicts that the age-grade system is stable? In the first place, the long-term behavior of the distribution is close to the distribution of the initial vector. If the society was able to tolerate the distribution of states when the system began, it will be able to tolerate distributions just as well in later generations. Even if the initial distribution vector was quite different from $\mathbf{w}$, there is still reason to conclude the system is stable. The age-grade system is most threatened if the proportion of older men in the lower grades continues to increase. The calculation of the limiting vector $\mathbf{w}$ shows that this proportion will remain, in the long run, under 30 percent. Whether the society can tolerate that high a proportion in the lowest age group is a question that can be decided only by more careful observation of the Gallas.

The discussion of the particular numbers obtained in Hoffmann's example is not central to a criticism of his approach. It should be pointed out that the data he used were not obtained by an actual observation or census of the Oromo people, but were chosen arbitrarily. Hoffmann wished to demonstrate how Markov chains could be used to study a problem of cultural stability. The entries of the transition matrix were chosen to represent

a not unreasonable situation for which no bias toward or away from stability was immediately apparent.

The critical assumption in Hoffmann's model is that the process is a Markov one—that is, that the transition matrix remains constant from generation to generation over many years. How reasonable is this assumption? Are the transition probabilities going to be the same for a generation of fathers that suffered through a drought or were decimated by illness or war as they are for a generation that has known plentiful harvest, good health, and peace? If one generation produces a set of males, most of whom enter the lowest grade at an advanced age, will the next generation try to adjust its birth rates to compensate for this condition?

Hoffmann's response to these criticisms is that there is value in the Markov chain approach even if there is no reason to believe that the transition matrix is constant [1971]: "If we are unwilling to postulate the invariance of the transition matrix, it is still possible to use the model as a decision procedure. The limiting vector of an observed transition matrix is readily calculated. Then one can state: 'This pattern of transitions is/is not compatible with the stability of the . . . system.'"

If we are willing, on the other hand, to postulate that the transition matrix remains constant, we can ask some important questions about Hoffmann's Markov chain model. In our early discussions about the age-grade system, we noted that it is quite conceivable that many males will enter the lowest grade before adolescence. This group is omitted entirely from Hoffmann's model. For completeness, he could have included a state S_0 corresponding to those who were initiated before the age of 13. If this state is included, then the transition matrix becomes a 4×4 array. If it is a regular matrix, then the theory of Chapter 10 still holds and a limiting vector $\mathbf{w}$ can be computed, although the calculations are more tedious.

In the exercises and suggested projects, you will be asked to explore further some possible modifications of this probabilistic model.

VI. Hans Hoffmann

Hans Hoffmann's work in anthropology ranged from field studies of Eskimo hunters and the cultures of the Amazon River basin to new theoretical developments in mathematical anthropology. He also conducted ethnographic research among a mental hospital population receiving new psychiatric drugs, and he served as a consultant on a project attempting a mathematical analysis of children's games. Hoffman developed a strong interest in maritime anthropology, studying how to apply mathematics to technical aspects of navigation and boat design.

Hoffmann was born in Koblenz, Germany, in 1929, but received his professional education in the United States. He did his undergraduate work at Cornell University where he was a mathematics major who devoted substantial time also to the fields of physics, astronomy, anthropology, and Chinese literature. He received his doctorate in anthropology from Yale University in 1957 for a thesis on cultural homogeneity among the Attawapsikat Cree. The field data for this study, which was supervised by an anthropologist and a psychologist, were gathered by Hoffmann and his wife Betty in James Bay, Canada.

Interest in the hunters of the northern forest led Hoffmann to the field work among the Eskimos and Crees in the mid-1950s. Among the unpublished material Hoffmann collected are reminiscences and ethnographic comments by his Cree informant gathered while Hoffmann observed his life in camp and accompanied him on hunting expeditions.

The Amazon River basin has also held a long-term fascination for Hoffmann. "This has led," he wrote, "to three field trips to the Shipibo of the Ucayali river in Eastern Peru. I am particularly interested in contemporary changes in technology and their effects on the

continuity of Amazonian cultures. This research is illuminated by an independent interest in lowland archeology. Several of my studies in mathematical anthropology are based on my Amazonian data. Conversely, mathematical culture theory has supplied an analytical framework for making field observations."

Hoffmann's developed pioneering mathematical models in theoretical anthropology. He published papers on deterministic, stochastic, and game theory models of cultural systems, material culture in a four-dimensional world, linear programming approaches to "cultural intensity," and an extensive survey of mathematical systems and their possible application to anthropological problems.

In a chapter for the *Biennial Review of Anthropology*, Hoffmann [1969] described the role of mathematical models:

> *Although human imagination is unbounded, our unaided ability to experience it is limited. Experiencing requires tools, and as these become developed, wider realms of imagination can be made one's own. We can imagine differences in the length of objects, but need a tool—the natural numbers—to experience them. We can imagine an infinity of numbers beyond the integers and their inverses, but need a tool—Cantor's diagonal proof—to experience their existence. For this reason, tools are the essence of culture; whether physical or mental, they permit man to experience wider ranges of his universe . . . . Mathematics is a tool that enables man to understand and control an immense number of events and processes in the physical world. Mathematics, in particular, is a tool that penetrates realms of imagination hopelessly beyond the experience of a toolless mind. Moreover, once mathematical tools have been developed, they often reverse their effect and enlarge not only one's experience but also one's imagination.*

Hoffmann taught briefly at the University of Oklahoma, University of Arkansas, and Cornell, before moving to the State University of New York at Binghamton in 1961. "From a long-range point of view," he said in the 1970s, "I do expect to return to empirical investigations when the mathematical issues that have intrigued me have been at least looked into. While mathematics is an exhilarating world to explore, it cannot really compete with the Amazon. Most likely I will combine a long-standing interest in the construction and sailing of small boats with that in the life of Indian traders on the Amazon tributaries. . . . I look forward to returning to the Amazon when the ages of my children make extensive field work feasible once more."

Photo courtesy of Hans Hoffmann

Hans Hoffmann and two of his children in early 1970s.

His colleague Daniel Strouthes noted that

Beginning in the early 1980s Hoffmann spent many summers in the Chesapeake Bay region, studying the waterways and the boats designed to fish them in the 19th century. He also purchased and sailed a Pacific-style multihull boat to broaden his knowledge of traditional Pacific voyaging. His maritime anthropology research resulted in his teaching 5 courses on the subject, and he thus single-handedly made Binghamton's maritime anthropology training the most extensive available at an inland U.S. university, and perhaps the most technologically oriented anywhere in the world.

Hoffmann's personal interests and love of anthropology were inseparable. He was an avid outdoorsman who combined his bicycling, hiking, cross-country skiing and gliding with his professional work. For example, his bicycling and hiking of the Erie Canal resulted in the introduction of a section on canals into his Technology and Material Culture course, and his long experience as a glider pilot led to his 1989 article Gliding in L3: Decisions, Decisions, *which used symbolic logic to describe decision-making processes of glider pilots. Hoffmann was an indefatigable researcher, possessed of an active mind that was always moving in new directions, and it was his students who were always the first and most significant beneficiaries of his work.*

Hans Hoffman died suddenly at his home in Vestal, New York on March 8, 1997.

EXERCISES

II. The Gadaa System

1. Meyer entered the age-grade system at the age of 15 and his son Frank was born when Meyer was 35 years old. Frank became the father of Michael at age 39. Michael's sons Eli and Alexander were born when he was aged 25 and 31, respectively. How old will Eli and Alexander be when they retire from the system?

2. Is it possible for a man to be born directly into the highest grade? Is it possible for him to be born after his "retirement" from the age-grade system?

III. A Deterministic Model

3. Prove Eq. (6) by mathematical induction.

4. What is the relation of A_{n+1} and A_1 if the average age of parenthood of the intervening n generations is less than 40 years? greater than 40 years?

5. How do the results of the deterministic model change if the length of time of an individual in the age-grade system is k years, instead of 40 years?

6. What would you estimate the life expectancy of an Ethiopian tribesman to have been in the 18th and 19th centuries? Is it likely that the average of parenthood could have been 40?

7. Note that the rules of the age grades do not allow a man to marry until he has been in the system for at least 16 years. What effect does this have on the average age of parenthood?

8. How would you modify the deterministic model to allow for the fact that some men have no sons, while others have more than one?

9. If the society is undergoing exponential or logistic population growth, will this affect the stability of the age-grade system?

10. Prins argued that "the functioning of the system of age-grades of the Galla . . . requires birth regulation as one of its basic institutional elements."

(a) Show that the results of Hoffmann's deterministic model indicate that this is *not* a necessary condition for stability.

(b) Prins claimed that restricting procreation to a man's last 12 years in the system would be the ideal way to achieve stability. In what sense, if any, is this true?

IV. A Probabilistic Model

11. Using Hoffmann's transition matrix P, calculate $\mathbf{p}^{(1)}$, $\mathbf{p}^{(2)}$, and $\mathbf{p}^{(3)}$ if

(a) $\mathbf{p}^{(0)} = (1, 0, 0)$

(b) $\mathbf{p}^{(0)} = (1/3, 1/3, 1/3)$

(c) $\mathbf{p}^{(0)} = (0, 3/4, 1/4)$

(d) $\mathbf{p}^{(0)} = (p, q, 1 - p - q)$

12. Repeat Exercise 11 with the transition matrix

$$p = \begin{array}{c} \\ S_1 \\ S_2 \\ S_3 \end{array} \begin{array}{c} \begin{array}{ccc} S_1 & S_2 & S_3 \end{array} \\ \begin{pmatrix} .6 & .3 & .41 \\ .1 & .8 & .1 \\ .1 & .2 & .7 \end{pmatrix} \end{array}$$

13. Repeat Exercise 11 if every entry in the transition matrix is 1/3.

14. Find the unique fixed point stochastic vector for

(a) the matrix P of Exercise 12,

(b) the matrix of Exercise 13.

Are these stable systems?

15. Add a fourth state S_0 for those who entered the system before age 13, assume that $\mathbf{p}^{(0)} = (.1, .2, .3, .4)$ and that the transition matrix is

$$\begin{array}{c} \\ S_1 \\ S_2 \\ S_3 \\ S_4 \end{array} \begin{array}{c} \begin{array}{cccc} S_1 & S_2 & S_3 & S_4 \end{array} \\ \begin{pmatrix} .5 & .3 & .15 & .05 \\ .2 & .6 & .15 & .05 \\ 0 & .2 & .7 & .1 \\ 0 & .05 & .15 & .8 \end{pmatrix} \end{array}$$

(a) Compute $\mathbf{p}^{(1)}$ and $\mathbf{p}^{(2)}$

(b) Is the transition matrix regular? If so, find its unique fixed-point stochastic vector. Does it give rise to a stable system?

16. Is it conceivable that the transition matrix might not be regular? What are the consequences of this for the model?

SUGGESTED PROJECTS

1. What are the effects on the stability of the age-grade system if one or more of the following modifications are made?

(a) A son enters the system at his father's death if his father dies before retirement.

(b) The eldest son enters exactly 40 years after his father does, but younger sons must wait until they are the same age as their older brother was when he was initiated.

(c) A son enters 40 years after his father does *or* at the age of 10, whichever event takes place later; thus, no one enters before the age of 10.

Can the deterministic and probabilistic models of this chapter be changed to incorporate these variations? Are new models necessary?

2. Discuss the practical problems of determining the entries of the transition matrix from observations and census data. Can you determine, in the absence of such information, bounds for the sizes of the entries of the transition matrix? Are all transition matrices whose entries satisfy these bounds necessarily regular?

3. Prins discusses age-grade systems among the Kipsigis and Kikuyus of Kenya as well as the Oromos of Ethiopia. Develop mathematical models for these age-grade systems. Do these systems have problems of stability?

You can find a listing of references and suggestions for additional reading on the book's website, www.wiley.com/college/olinick

CHAPTER 13 Paired-Associate Learning

The mind is slow to unlearn what it has been long in learning.

—Seneca

I. The Learning Problem

The study of learning has been a basic concern of psychology for more than a century. There is a vast literature of books and articles developing different theories or presenting the results of learning experiments involving human and animal subjects. In his chapter, we will examine a very simple model of a particular kind of learning situation. The model was developed by Gordon Bower (and independently by R. R. Bush and F. A. Mosteller) around 1960 in the early days of mathematical learning theory. Mathematical model building for learning processes has had a rich and varied history during the past 50 years; the current "state of the art" has advanced quite far beyond the material we will study. Bower's work is worth examining for us because it shows how a number of predictions can be deduced from a simple model and because it illustrates an actual use of the absorbing Markov chains studied in Chapter 11.

Bower examined a learning problem exemplified by a task familiar to most students. Suppose you are studying for a vocabulary test in your Swahili course. You must be able to translate into Swahili a prescribed list of 25 English words. This learning situation demands the following:

1. You must learn the Swahili words. This includes proper pronunciation and spelling.
2. You must learn to match the correct Swahili word to the appropriate English word.

Bower's model of "paired-associate" learning (PAL) is concerned with the second task, the associative "hook-up" of the relevant responses to their appropriate stimulus members. Paired-associate learning was invented by Mary Whiton Calkins in 1894. Calkins (1863–1930) established the first research laboratory in psychology at a liberal arts college, Wellesley College. She was the first female president of both the American Psychological Association (APA) in 1905, and the American Philosophical Association in 1918.

To describe the learning situation more carefully, suppose that the experimenter determines a set of ordered pairs (s, r) where s is chosen from a finite set S, called the *stimulus set*, and r is selected from a finite collection R, labeled the *response set*. An element s is shown to a subject and she tries to give the corresponding r. The subject, in other words, is trying to find the appropriate value $f(x)$ of a function when she is told the domain value x.

In the example of the foreign language word list, the set S consists of 25 English words and the set R of 25 Swahili words. The function f for this example is one-to-one, but it need not have this property in general.

When the subject is first presented with a stimulus, she can only guess what the appropriate response is supposed to be. After she guesses, the experimenter tells her what the correct response should have been. This is the first point in the experiment at which learning may occur.

If there are K elements in the set S, then a *trial* is defined as one cycle of presentation of each of the K items, the order of appearance of the items being randomized over successive trials. After the first trial, the subject may either guess again or give the correct response because she has learned it from the result of a previous trial.

Each subject in the experiment responds to each element of S on the first trial. The order of presentation of the stimulus elements is scrambled, and the subject is asked again. This procedure is repeated until the subject responds correctly to all elements of S on two consecutive trials. It is then assumed that the subject has completed the learning task. In theory, the experiment for a single subject could last for indefinitely many trials, but in practice, the learning task is chosen so that it is completed by all subjects by the 20th trial.

Data is collected from the experiment by recording each correct response by a 0 and each incorrect response by a 1. If we fix our attention on a single element of S, then the subject generates a sequence

$$x_1, x_2, \ldots, x_n, \ldots$$

of 0s and 1s. To repeat, the number x_i is 0 if the subject made the correct response to the particular stimulus s when presented with it on the ith trial, and it is a 1 if there was an incorrect response on the ith trial.

> *"Stripped to its barest essentials," Bower [1961] explains, "the job for a theory of PAL is to describe and account for the general characteristics of these sequences. The best job of description, of course, would be to reproduce the original sequences. Theories, as economic abstractions, do not perform this task but they can provide general descriptions (e.g., the trial number of the second success) about a sample of sequences allegedly generated under the same process laws. Obviously models that deliver predictions about many different aspects of such sequences are preferable to less tractable models, since each prediction provides an opportunity to test the adequacy of the model. In turn, the number of predictions derivable in closed form from a model reflects to a large extent the simplicity of the assumptions used to represent the process under consideration. The assumptions of the model to be presented appear to achieve almost maximal simplicity for a model about learning; accordingly, it is possible to derive in closed form an extensive number of predictions (theorems) referring to properties of the response sequences obtained from the learning subject."*

II. The Model

A. Axioms of the Model

Bower's model assumes at the start that each stimulus item in the list of paired associates may be represented by exactly one element from a set S and that the correct response to that stimulus item becomes associated in an all-or-none fashion. He also assumes that the

subject knows the elements of the response set before the experiment begins; thus, the model can concentrate on the second associative aspect of the learning problem as discussed in Section I.

The process of associating elements in S with elements in R is governed, according to the model, by five basic axioms:

AXIOM 1 On each presentation of an item from the set S, only two states, C and U, are possible for the subject with respect to this item. If the subject is in state C for that item, then she knows the correct response and will give it. If the subject is in state U for that item, then she does not know the proper response and she guesses an element from the permitted set of responses.

If the subject knows the proper response, we say she is in the *conditioned state* C; otherwise, she is in the *unconditioned state* U.

AXIOM 2 At the beginning of the experiment, the subject is in state U for each item in S.

AXIOM 3 The state C is an absorbing state; if an item has become conditioned, then continued study of the same correct response will ensure that the item remains conditioned.

AXIOM 4 If the subject is in state U immediately preceding any trial, then the transition probability moving from U to C on the next trial is a positive constant c, which is the same for each trial, each item, and each subject.

AXIOM 5 If the subject is in state U, then the probability that she guesses the correct response is $\frac{1}{N}$ where N is the number of elements in the set R.

These axioms contain a number of simplifying assumptions about how humans learn, some of which may seem startlingly naive to you. Does it indeed seem reasonable to assume that there is a transition probability c that does not vary from person to person? Even for a single subject, is it not likely that the chances of becoming conditioned to an item would depend on the nature of the item?

The assumption (Axiom 5) that a subject in state U chooses from the response set with equiprobability also seems suspect. While she may not have learned, for example, to associate the Swahili word *twiga* with the English word *giraffe*, the subject may know that the correct response is one of three or four Swahili words from the total list of 25. In other words, she may be guessing at random not from the set R, but from some smaller subset R'. Also, if she has become conditioned to associate, say, the Swahili *simba* with the English *lion*, then she will not respond with *simba* when asked to translate *giraffe*. Thus, as more items become conditioned, the probability of a correct guess should increase.

Several responses to these criticisms of the axioms are possible. We shall see how Bower designed a specific learning situation in which some of these objections cannot arise. More important, however, every mathematical model contains simplifying assumptions. Whether these simplifications are reasonable ones to make can only be judged after the predictions of the model are compared with observations of the real-world system that is being modeled. In Section III, we will make these comparisons. Let's examine now what we can deduce from the set of axioms.

The words "state" and "transition probabilities" are used in the statements of the axioms deliberately. They indicate that we mean to use Markov chains to study the paired-associate learning situation. Axioms 1, 3, and 4 assert that the learning process is an absorbing Markov chain. From these axioms, we can construct the transition matrix, which describes the movement from state to state in successive trials for a given item.

The transition matrix has the form

$$P = \begin{array}{c} \\ C \\ U \end{array} \begin{array}{c} \begin{array}{cc} C & U \end{array} \\ \begin{pmatrix} 1 & 0 \\ c & 1-c \end{pmatrix} \end{array}$$

where the rows give the state prior to the start of one trial and the columns give the state prior to the start of the next trial.

From Axiom 2, the initial probability vector is

$$p^{(0)} = (0, 1)$$

because the subject does not know at the outset how the items have been associated.

The model has a single parameter, c, which is the likelihood that an unconditioned item will become conditioned as the result of a reinforced trial (evoking the correct response). The effect of successive reinforced trials is to provide repeated opportunities for the item to become conditioned.

From a subject's response to the stimulus on a particular trial, we cannot always assert which state is being occupied. A correct response is certain if the subject is in state C, but it is also possible that the subject was in state U and made a lucky guess. If, however, the subject makes an incorrect response on the trial, then the subject must have been in the unconditioned case. This fact makes it possible to prove a number of theorems about the model.

B. Predictions of the Model

Bower derives a large number of predictions about the learning process from his Markov chain model. We will list some of them here and give proofs of a few.

It is possible to deduce from the model the following:

A. The average number of trials before learning the item

B. The probability q_n, of an error on the nth trial—in fact, this is really an infinite number of predictions: $q_1, q_2, \ldots$

C. The average number, u_1 of errors before learning an item

D. The probability of a run of k consecutive errors $(k = 1, 2, 3, \ldots)$ that start on the nth trial for $n = 1, 2, 3, \ldots)$

E. The expected value of r_k, where r_k represents the number of error runs of length k (k consecutive errors followed by a correct response)

F. The expected value of R where $R = r_1 + r_2 + \cdots$, the total number of error runs

G. The probability distribution of T, the total number of errors on each item—that is, we deduce $\Pr(T = k)$ for $k = 0, 1, 2, \ldots$ (note that the expected value of T is $EV(T) = u_1$)

H. The expected value of $c_{k,n}$, where $c_{k,n}$ is the number of times that an error on trial n is followed by an error k trials later

I. The expected values c_k of the "autocorrelations" of x_n and x_{n+k} over all trials of the experiment—that is,

$$c_k = EV\left[\sum_{n=1}^{\infty} x_n x_n + k\right] = \sum_{n=1}^{\infty} EV(c_{k,n}).$$

For instance, c_2 is the average number of times errors occur two trials apart.

J. The average number of alternations of successes and failures

K. The average number of errors before the kth success, $k = 1, 2, \ldots$

L. The proportion of items for which there are no errors following the kth success for $k = 1, 2, \ldots$

M. The probability distribution for the number of errors between the kth and the $(k + m)$th success, for all positive integers k and m

N. The probability distribution of the number of successes between adjacent errors

C. Deriving the Predictions

In this section, we will show how some of the predictions (A)–(N) may be deduced from the model.

We begin by writing the transition *matrix P* in the standard form (Chapter 11, IV):

$$\begin{array}{l} \\ \text{Absorbing States} \\ \text{Transient States} \end{array} \begin{array}{c} \begin{array}{cc} \text{Absorbing} & \text{Transient} \\ \text{States} & \text{States} \end{array} \\ \left(\begin{array}{c|c} I & 0 \\ \hline R & Q \end{array}\right) \end{array}$$

In this case, we obtain

$$P = \begin{array}{c} \\ C \\ U \end{array} \begin{array}{c} \begin{array}{cc} C & U \end{array} \\ \left(\begin{array}{c|c} 1 & 0 \\ \hline c & 1-c \end{array}\right) \end{array}$$

so that Q is the 1×1 matrix $(1 - c)$. Thus, $I - Q = 1 - (1 - c) = c$. Hence, the fundamental matrix is

$$N = (I - Q)^{-1} = \frac{1}{c}$$

The following conclusions are immediate:

1. Since every such Markov process eventually reaches an absorbing state, every subject will eventually learn every item;
2. The average number of trials per item that a subject is in the unconditioned state is $\frac{1}{c}$.

In this way, we have derived prediction (A):

THEOREM 1 The average number of trials before learning the item will be $\frac{1}{c}$.

Our next task will be to show that the probability q_n of an error on the nth trial of an item is given by

$$q_n = (1-c)^{n-1}\left(1-\frac{1}{N}\right)$$

This result is true essentially for two reasons. First, the subject must fail to be conditioned on each of the first $n-1$ trials. On each trial, this happens with probability $1-c$. Second, the subject must guess incorrectly on the nth trial; according to Axiom 5, this happens with probability $\left(1-\frac{1}{N}\right)$.

In order to give a more rigorous proof, we need to introduce a little extra notation.

We define C_n to be the event that the subject is in state C immediately prior to the response on the nth trial and U_n the event that she is in state U at that time. These are mutually exclusive events, so we have

$$\Pr(C_n) + \Pr(U_n) = 1$$

Axiom 3 gives us the result that $\Pr(C_1) = 0$ and $\Pr(U_1) = 1$. The axioms of the model also give us

$$\Pr(x_n = 1|C_n) = 0 \text{ and } \Pr(x_n = 1|U_n) = 1 - \frac{1}{N} \tag{1}$$

In the notation of Markov chains we have $\mathbf{p}^{(0)} = (\Pr(C_1), \Pr(U_1))$

$$\mathbf{p}^{(1)} = (\Pr(C_2), \Pr(U_2)) = \mathbf{p}^{(0)}P$$

$$\ldots$$

$$\mathbf{p}^{(n-1)} = (\Pr(C_n), \Pr(U_n)) = \mathbf{p}^{(0)}P^{n-1}$$

$$\mathbf{p}^{(n)} = (\Pr(C_{n+1}), \Pr(U_{n+1})) = \mathbf{p}^{(0)}P^{n}$$

$$\ldots$$

Ordinary matrix multiplication and simplification shows that

$$p^2 = \begin{pmatrix} 1 & 0 \\ c & 1-c \end{pmatrix}\begin{pmatrix} 1 & 0 \\ c & 1-c \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ c+c(1-c) & (1-c)^2 \end{pmatrix}$$
$$= \begin{pmatrix} 1 & 0 \\ 1-(1-c)^2 & (1-c)^2 \end{pmatrix}$$

A simple induction argument then leads to the conclusion that

$$P^k = \begin{pmatrix} 1 & 0 \\ 1-(1-c)^k & (1-c)^k \end{pmatrix} \quad k = 1, 2, \ldots \tag{2}$$

With k set equal to $n-1$, we obtain

$$(\Pr(C_n), \Pr(U_n)) = \mathbf{p}^{(0)}P^{n-1} = (0,1)P^{n-1} = (1-(1-c)^{n-1}, (1-c)^{n-1}) \tag{3}$$

so that

$$\Pr(C_n) = 1-(1-c)^{n-1} \text{ and } \Pr(U_n) = (1-c)^{n-1} \tag{4}$$

We have enough machinery now to establish prediction (B). ◇

THEOREM 2 The probability of an error on the nth trial is given by

$$(1-c)^{n-1}\left(1-\frac{1}{N}\right)$$

Proof of Theorem 2 The probability of an error on the nth trial can be represented as $\Pr(x_n = 1)$. Since the subject was in one of the mutually exclusive states C or U before the nth trial began, we have

$$\Pr(x_n = 1) = \Pr(x_n = 1 \cap U_n) + \Pr(x_n = 1 \cap C_n)$$
$$= (x_n = 1 | U_n)\Pr(U_n) + \Pr(x_n = 1 | C_n)\Pr(C_n)$$

where the second equality comes from elementary results about conditional probabilities (Chapter 10, II).

Now we make use of Eqs. (1) and (4) to write

$$\Pr(x_n = 1) = \left(1-\frac{1}{N}\right)(1-c)^{n-1} + 0(1-(1-c)^{n-1})$$

the desired result.

Our next deduction from the axioms gives a result that is useful in determining the value of the parameter c in experimental situations. ◇

THEOREM 3 The expected total number of errors, u_1, before learning an item is

$$\frac{\left(1-\frac{1}{N}\right)}{c}$$

To obtain this result, we first introduce a new random variable. Let T_M denote the total number of errors made by a subject on a particular item in the first M trials. Since $x_n = 1$ if the subject makes an error on the nth trial and is 0 otherwise, we have

$$T_M = x_1 + x_2 + \cdots + x_M$$

in other words, T_M is itself the sum of M random variables. We want to compute the expected value of T_M. It is clear that the expected value of T_M will be related to the expected values of the x_is. It turns out that this relationship is a particularly simple one. ◇

LEMMA Let R_1 and R_2 be two random variables defined on the same finite set E, which has probability measure Pr. Then $EV(R_1 + R_2) = EV(R_1) + EV(R_2)$.

Proof of Lemma

$$\begin{aligned} EV(R_1 + R_2) &= \sum_{x \text{ in } E} (R_1 + R_2)(x)\Pr(x) \\ &= \sum_{x \text{ in } E} (R_1(x) + R_2(x))\Pr(x) \\ &= \sum_{x \text{ in } E} [R_1(x)\Pr(x) + R_2(x)\Pr(x)] \\ &= \sum_{x \text{ in } E} R_1(x)\Pr(x) + \sum_{x \text{ in } E} R_2(x)\Pr(x) \\ &= EV(R_1) + EV(R_2) \end{aligned}$$

An easy induction argument establishes the corollary. ◇

Corollary The expected value of a finite sum of random variables is the sum of their expected values. ◇

We can use the corollary to begin to compute the expected value of T_M:

$$\begin{aligned} EV(T_M) &= EV(x_1 + x_2 + \cdots + x_M) \\ &= EV(x_1) + EV(x_2) + \cdots + EV(x_M) \end{aligned}$$

The computation of the expected value of an x_i is easy:

$$\begin{aligned} EV(x_i) &= 1\Pr(x_i = 1) + 0\Pr(x_i = 0) \\ &= \Pr(x_i = 1) \end{aligned}$$

$= \left(1 - \frac{1}{N}\right)(1-c)^{n-1}$, by Theorem 2.

Putting these results together, we have

$$\begin{aligned} EV(T_M) &= \left(1 - \frac{1}{N}\right)(1-c)^0 + \left(1 - \frac{1}{N}\right)(1-c)^1 + \cdots + \left(1 - \frac{1}{N}\right)(1-c)^{M-1} \\ &= \left(1 - \frac{1}{N}\right)\left[(1-c)^0 + (1-c)^1 + \cdots + (1-c)^{M-1}\right] \end{aligned}$$

Now the expression in square brackets is the sum of the first M terms of a geometric progression with first term 1 and common ratio $(1-c)$. Thus,

$$EV(T_M) = \left(1 - \frac{1}{N}\right)\left[\frac{1-(1-c)^M}{1-(1-c)}\right] = \left(1 - \frac{1}{N}\right)\left[\frac{1-(1-c)^M}{c}\right] \tag{5}$$

We find the expected number of total errors that will be made before an item is learned by letting $M \to \infty$ in Eq. (5) so that $(1-c)^M \to 0$. This establishes the statement of Theorem 3.

Finally, we will derive some results on the extent to which an error on a given trial tends to be followed by an error some number of trials later. Let $c_{k,n}$ denote the product

$$c_{k,n} = x_n \cdot x_{n+k}$$

This product has value 1 exactly when the subject makes errors on the nth trial and on the $(n+k)$th trial; otherwise it is 0. The expected value of $c_{k,n}$ is then given by

$$\begin{aligned} EV(c_{k,n}) &= \Pr(x_{n+k} = 1 \cap x_n = 1) \\ &= \Pr(x_{n+k} = 1 | x_n = 1)\Pr(x_n = 1) \\ &= \Pr(x_{n+k} = 1 | x_n = 1)\left(1 - \frac{1}{N}\right)(1-c)^{n-1} \end{aligned}$$

To find the conditional probability in this equation, we examine how an error occurs on the $(n+k)$th trial. It must be the case that conditioning fails during each of the k trials after the nth one and also that the subject guesses incorrectly on the $(n+k)$th trial. Thus, we have

$$\Pr(x_{n+k} = 1 | x_n = 1) = (1-c)^k\left(1 - \frac{1}{N}\right)$$

We find then that

$$EV(c_{k,n}) = (1-c)^k\left(1 - \frac{1}{N}\right)\left(1 - \frac{1}{N}\right)(1-c)^{n-1}$$

The average value of the "autocorrelation" of x_n and x_{n+k} over all trials is

$$c_k = EV\left[\sum_{n=1}^{\infty} x_n x_{n+k}\right] = \sum_{n=1}^{\infty} EV(c_{k,n}) = \sum_{n=1}^{\infty}\left(1 - \frac{1}{N}\right)^2 (1-c)^k (1-c)^{n-1}$$

so that

$$c_k = \left(1 - \frac{1}{N}\right)^2 (1-c)^k \sum_{n=1}^{\infty} (1-c)^{n-1}$$

This last sum is an infinite geometric progression with initial term 1 and common ratio $(1-c)$. The sum equals $\dfrac{1}{(1-(1-c))} = \dfrac{1}{c}$. Hence, we have

$$c_k = \frac{(1-c)^k \left(1 - \frac{1}{N}\right)^2}{c}$$

This gives us the predictions (H) and (I) promised in the preceding section. In this derivation, we have used infinite sums rather recklessly. The arguments can be made rigorous by restricting ourselves first to finite sums and then employing a limiting process. The procedure is much the same as in the proof of Theorem 3. The reader is encouraged to supply the details.

By arguments similar to the ones of this section (some easier, others more complex), we can deduce formulas for the other predictions (A)–(N) of the model. You will do this in the Exercises.

Now that we have good number of predictions of how subjects would behave in a paired-associate learning situation if the axioms are correct, we can turn to the task of comparing them with real-world observations.

III. Testing the Model

Bower compared the predictions of his model with the results obtained in an experiment involving 29 subjects. He presented each subject with a list of 10 pairs of consonant letters; these pairs constituted the stimuli. The subject had to learn to associate each pair with either the integer 1 or the integer 2. For each subject, five of the pairs were selected at random to be associated with 1; the correct response for the other five pairs then was 2. The experiment continued until the subject was able to complete two consecutive cycles of all 10 pairs. The letters were written on cards and the cards were shuffled between trials to randomize the order of presentation of the stimuli.

Since "1" is the correct response to five different stimuli, the subject cannot discard any element in the response set even after becoming conditioned to some of the stimulus items. Also, as there are only two possible responses, the subject has to guess—when guessing is necessary—from the full response set. Note how the design of the experiment deals with some of the objections we raised earlier about the simplicity of the axioms.

Bower observed that the average number of errors per item made by his subjects was 1.45. Since $N = 2$ in this experiment, Theorem 2 gives a predicted average of $1/(2c)$ errors. Equating the predicted value with observed one gives $c = 1/2.9 = .345$. This estimate of c will be fixed throughout the remaining discussion of the data.

A major feature of interest to psychologists in experiments like Bower's is the "learning curve." This is a graph of the percentage of incorrect responses as a function of the number of trials. To obtain the observed learning curve, we plot the proportion of wrong responses versus the number of trials and then connect these points with a smooth curve. The theoretical learning curve is derived from the model and is the graph of q_n as a function of n.

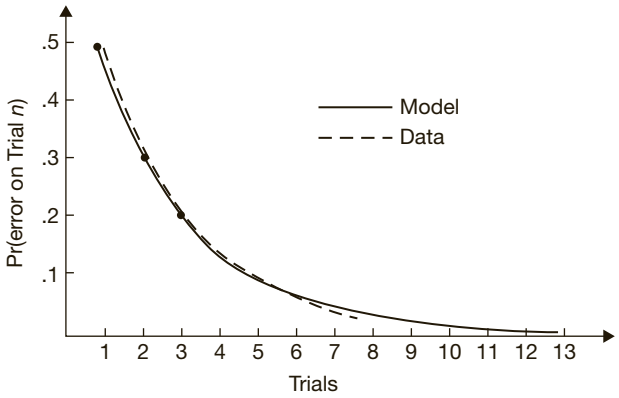


FIGURE 13.1 The probability of an incorrect response over successive trials of the experiment. Taken from Bower (1961), with permission.

Fig. 13.1 is from Bower's paper [1961] and it shows how closely the predicted learning curve fits the observed one.

Another graphic example of how well the model fits the observed data is provided by the distribution of T, the total number of errors per item. Bower's Markov chain model predicts that

$$\Pr(T=k)=\begin{cases}\dfrac{b}{N} & \text{for } k=0\\[2ex] \dfrac{b(1-b)^k}{1-c} & \text{for } k\geq 1\end{cases}$$

where b is the constant

$$b=\frac{c}{1-\dfrac{1-c}{n}}$$

For $c=.345$ and $N=2$, we have $b=.513$, so the model predicts

$$\Pr(T=k)=\begin{cases}.256 & \text{for } k=0\\[2ex] \dfrac{.513(.487)^k}{.655} & \text{for } k\geq 1\end{cases}$$

The graphs of predicted and d observed distributions of T are shown in Fig. 13.2.

Bower made a number of other comparisons of the data collected from his actual experiment with the predictions of his model. Some of these are collected in Table 13.1.

Since the observed value of the average number of errors per item was used to calculate c, we automatically get perfect agreement for the first statistic. It is rather remarkable that the other 21 pairs of numbers are so close together in value.

One final comparison of Bower's model with experimentally observed data may be of interest. According to the axioms of the model, if the subject makes a mistake on the nth trial, then the item was not conditioned prior to the start of that trial. The subject was in state U when the nth trial began, just as when the first trial started. The model asserts that the degree

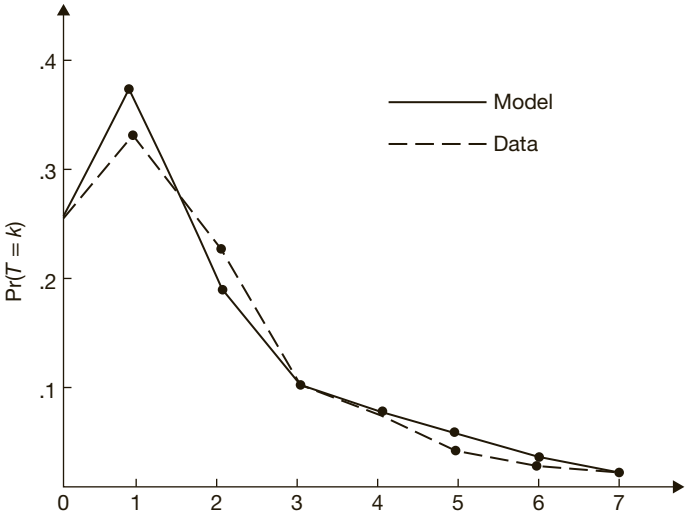


FIGURE 13.2 Distribution of T, the total number of errors per item.

Table 13.1 Comparison of model's prediction and observed data

Statistic	Prediction of model	Observed data
1. Average number of errors/item	1.45	1.45
2. Standard deviation (SD) of (1)	1.44	1.37
3. Average number of errors before first success	.749	.785
4. SD of (3)	.98	1.08
5. Average number of errors between first and second success	.361	.350
6. SD of (5)	.76	.72
7. Average number of errors before second success	1.11	1.13
8. SD of (7)	1.10	1.01
9. Average number of successes between errors	.488	.540
10. SD of (9)	.72	.83
11. Average trial of last error	2.18	2.33
12. SD of (11)	2.40	2.47
13. Total error runs	.973	.966
14. Error runs of length 1	.655	.645
15. Error runs of length 2	.215	.221
16. Error runs of length 3	.070	.058
17. Error runs of length 4	.023	.024
Autocorrelation of errors:		
18. One trial apart (c_1)	.479	.486
19. Two trials apart (c_2)	.310	.292
20. Three trials apart (c_3)	.201	.187
21. Alternations of success and failure	1.45	1.143
22. Probability of a success following an error	.672	.666

of the subject's associative connection with that item and the correct response has not effectively changed since the experiment started. In terms of predicting the subject's future behavior on this item, we get the same results whether or not we neglect the first $n - 1$ trials.

In particular, we may consider the average number of errors that follow an error on trial n. According to Bower's model, this number is the constant $u_1(1 - c) = (1 - c)\dfrac{1 - \frac{1}{N}}{c}$, which is independent of n. This prediction is in sharp contrast to the prediction of the "linear model" of learning. The linear model predicts that the number of errors expected following an error on trial n should be a decreasing function of n, since associative strength is assumed to increase steadily with the number of preceding reinforced trials.

To test which, if either, prediction was correct, Bower used the data from the 29 subjects of the experiment we have described, along with data from 47 other subjects involve in similar learning experiments. The results, shown in Fig. 13.3, show that Bower's model is much closer to the observed data.

In summary, then, a simple model seems to predict quite well the results of learning in a simple paired-associated task. As Bower [1961] concludes,

> *"The fact that the . . . model gives an adequate quantitative account of these paired-associate data satisfies one important requisite of a scientific theory, that of being close to the data. If, in addition, the theory is mathematically tractable in that numerous consequences are easily derived in closed form, then indeed we are in a fortunate position. The main task of this paper has been to show that the . . . model is mathematically tractable. . . . This property of the model is due to the extreme simplicity of its assumptions about the association process. One might effectively argue that the present model nearly achieves the absolute minimum in assumptions for a workable theory of learning.*
>
> *"Once one has demonstrated the predictive validity of a model for a limited class of experimental situations, there remains the task of characterizing more generally those experimental arrangements to which the model may be expected to apply. . . . We explicitly restricted the model to the S–R association process and have used simplified experimental situations in which response learning was precluded. Within this restricted domain of paired-associate learning, the model has proved extremely useful in investigating the effects on learning of variations in the number of response alternatives and in the reinforcement*

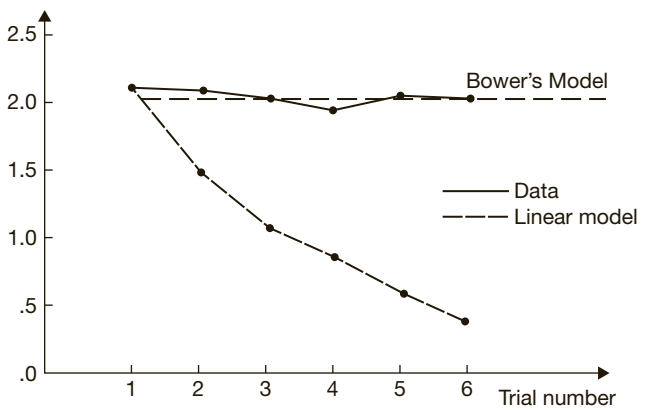


FIGURE 13.3 Average number of errors following an error on trial n.

conditions prevailing during learning. . . . Ultimately, one would like to have a set of combination axioms whereby the assumptions about S–R association and response learning may be combined for predicting results in those experimental situations involving the concurrent operation of these two processes."

IV. Historical and Biographical Notes

A. Mary Whiton Calkins

The story of Mary Whiton Calkins is one of triumph over American sexism, especially prejudice against women in higher education. Denied a doctoral degree from Harvard University because of her gender, Calkins went on to a distinguished career in psychology and philosophy, ultimately serving as the first female president of both the American Psychological Association and the American Philosophical Association.

Calkins was born March 30, 1863, in Hartford Connecticut, the oldest of five children. Calkins's parents reared their family with an international and cosmopolitan focus, speaking only German to them when they were very young and taking frequent trips to Europe so their children would also become fluent in French. At age 17, Calkins moved to Massachusetts when her father, a prominent Protestant clergyman, accepted a position in Newton.

Mary Calkins entered Smith College with sophomore standing, took a year off school after her sister died to study on her own, and returned to receive her undergraduate degree in 1885. After graduation, she spent 16 months traveling abroad with family and then returned to Massachusetts where she worked for three years as a tutor in Greek at Wellesley College. At this time, the field of psychology was emerging from an area within philosophy to an independent discipline of its own, a development accelerated by William James's monumental *Principles of Psychology*, first published in 1890. Wellesley's philosophy department chair asked Calkins whether she would be willing to develop and teach a new course in psychology. She agreed, provided she was given a year to learn more about the subject.

Calkins decided she would be best off studying at Harvard with James and with Josiah Royce, the eminent American philosopher. They both agreed to take her on as a graduate student, but Harvard President Charles William Eliot initially refused to let her attend, as he did not believe men and women should study together in the same room. Royce, James, Wellesley College President Helen Shafer, and Wolcott Calkins, Mary's father, all appealed the decision and Eliot relented. Mary could take regular classes along with the men, but Eliot stipulated that she could not officially register as a student.

When Mary showed up for her classes with James, she discovered that none of the men who had previously registered showed up. As she reported in an autobiographical essay,

I began the serious study of psychology with William James. Most unhappily for them and most fortunately for me the other members of his seminary in psychology dropped away in the early weeks of the fall of 1890; and James and I were left . . . quite literally at either side of a library fire. The Principles of Psychology *was warm from the press; and my absorbed study of those brilliant, erudite, and provocative volumes, as interpreted by their writer, was my introduction to psychology. What I gained from the written page, and even more from* tête-à-tête *discussion was, it seems to me as I look back upon it, beyond all else, a vivid sense of the concreteness of psychology and of the immediate reality of "finite individual minds" with their "thoughts and feelings."*

Calkins completed a Ph.D. under the tutelage of her Harvard mentors, who staged an unofficial thesis defense wherein she was examined by five faculty luminaries in psychology and philosophy. They found her work worthy of a degree with honors, but the Harvard administration refused to confer a doctorate on a woman. Although she later received honorary degrees from institutions such as Columbia and Smith, Harvard never awarded her the recognition her work deserved.

To deepen her background in experiment psychology, Calkins also worked with Clark University's Edmund Sanford. Sanford also provided some assistance to Calkins when she set up a psychology laboratory at Wellesley College, the first such lab at a liberal arts college. Calkins pursued an active teaching and research career. She created and developed paired-associate learning as a technique to study human memory. In 1896, she published a description of her experiments on the task of associating colors with numbers, examining the effect of such factors as the vividness of exposure or the length of time a color was exposed to the subject. She also devoted considerable effort to the study of the concept *self*, ultimately concluding that it could be precisely defined but was "a totality, a one of many characters . . . a unique being in the sense that I am I and you are you. . . ." Sigmund Freud cited Calkins's work on dreams as an influence on his theory of dreams.

Mary Whiton Calkins (right) and her psychology laboratory at Smith College.

Calkins was nationally and internationally recognized as an outstanding psychologist and philosopher. In a 1903 survey, she was named in the top dozen of American psychologists. She published more than 100 articles and four more extended volumes. Among her writings in philosophy, two book stand out: *The Good Man and the Good: An Introduction to Ethics* and *The Persistent Problems of Philosophy: An Introduction to Metaphysics through the Study of Modern Systems*. Calkins retired from Wellesley in 1929. She died on February 26, 1930.

Spurred in part by her own experience as a victim of gender bias, Calkins decried prejudice whether she found it in published research or observed it in every day life. She was active in the Consumer's League and the American Civil Liberties Union and advocated for equality for women. In 1902, Calkins was offered a Radcliffe Ph.D. degree, which she declined on principle. Believing that work done at Harvard should be recognized by a Harvard degree regardless of whether the recipient was a man or woman, she declined the

Radcliffe offer. Throughout her career, Calkins registered opposition to differentiation between the sexes based on the erroneous assumption of inherent differences in mental abilities.

A major political issue for women during Calkin's lifetime was suffrage, the right to vote, which was not extended to all women in the United States until the passage of the 19th Amendment to the Constitution. Speaking before the National Suffrage Convention in Baltimore, Calkins said

> *"The student trained to reach decisions in the light of logic and of history will be disposed to recognize that, in a democratic country, governed as this is by the suffrage of its citizens, and given over as this is to the principle and practice of educating women, a distinction based on difference of sex is artificial and illogical."*

B. Gordon H. Bower

Gordon Howard Bower is a distinguished cognitive psychologist specializing in experimental studies of human memory, language comprehension, emotion, and behavior modification. He spent most of his professional career at Stanford University from which he retired as Albert Ray Lang Professor Emeritus of Psychology in 2005.

Born was born in the small town of Scio, Ohio, during the Great Depression, in 1932. Inspired by the movie *The Lou Gehrig Story*, Gordon resolved at the age of 8 to become a professional baseball player. By 11, he played on local semi-professional baseball teams. As a young man, he helped out in his father's general store and worked on several local farms. A talented baseball and basketball player, Bower received an athletic scholarship to attend Western Reserve University. He gave up the opportunity for a professional baseball career so that he could pursue his interest in psychology.

After graduating from Western Reserve in 1954, Bower spent a year at the University of Minnesota studying the philosophy of science under a Woodrow Wilson Fellowship. He left the Midwest to continue his graduate program at Yale University, where he was awarded his doctorate in psychology in 1959.

Photo courtesy of Gordon Bower

Gordon H. Bower

In 1957, Bower won a fellowship to attend a summer workshop on mathematical learning theory at Stanford. There he met many of the contributors to the burgeoning field of mathematical psychology. While attending that workshop, Bower so impressed the Stanford faculty that he was offered a job before he had finished his Ph.D. thesis at Yale.

Bower's research and teaching interests have centered on conditioning, learning, human memory, mathematical models, and computer simulation of memory processes. He has written nearly 250 technical articles and four books.

The quality and significance of Bower's research has been recognized by many honors, including elections to the prestigious National Academy of Sciences and the American Academy of Arts and Sciences. In 2007, President George W. Bush awarded Bower the National Medal of Science, the nation's highest science award. The White House cited Bower "for his unparalleled contributions to cognitive and mathematical psychology, for his lucid analyses of remembering and reasoning and for his important service to psychology and to American science."

As a teacher, Bower urged his students to be active readers of the research literature, imagining ways they can contribute to the field. He encouraged students to follow their own interests, but to try to be at the forefront of new developments.

"Students are my treasures," Bower says. "I get great satisfaction from their accomplishments."

The preface to a *festschrift* in Bower's honor fittingly concludes:

> *Gordon never fulfilled his early dream of pitching a no-hitter at Yankee stadium. . . . However, in his chosen career of psychology, where he went up to bat time after time against a broad and diverse lineup of the most challenging problems in learning and memory, Gordon hit a string of home runs worthy of his childhood idol, Lou Gehrig.*

Photo by Christy Bowe. Courtesy of the George W. Bush Presidential Library & Museum

President George W. Bush presents the National Medal of Science to Gordon Bower.

EXERCISES

1. Let $S_M = 1 + r + r^2 + \cdots + r^{M-1}$ where r is a real number and M is a positive integer.

(a) Show that $S_M - r S_M = 1 - r^M$ so that $S_M = \frac{(1 - r^M)}{(1 - r)}$, if $r \neq 1$.

(b) If $|r| < 1$, show that $S = \lim_{M \to \infty} S_M = \frac{1}{1 - r}$; in this case, we say $S = 1 + r + r^2 \cdots +$

(c) Find the value of $a + ar + ar^2 + \cdots$ where $|r| < 1$ and a is a constant.

(d) What happens to these sums in the cases

(i) $|r| > 1$?

(ii) $r = 1$?

(iii) $r = -1$?

2. What are the predictions of Bower's model if

(a) $c = 0$?

(b) $c = 1$?

Do either of these cases have relevance for human learning?

3. Investigate the consequence of Bower's model if the initial vector is $p^{(0)} = \left(\frac{1}{2}, \frac{1}{2}\right)$. How might this initial vector occur in an experimental situation?

4. Prove, by induction on k, that the kth power of the transition matrix of the Bower model can be written in the form

$$P^k = \begin{pmatrix} 1 & 0 \\ 1-(1-c)^k & (1-c)^k \end{pmatrix}$$

for all $k = 1, 2, 3, \ldots$ [*Hint*. It is sufficient to check only the entries in the second column; why?]

5. Prove the corollary to the Lemma of Section II.

6. A mathematically rigorous derivation of the formula for c_k can be given. Define

$$c_k^{(m)} = EV\left[\sum_{n=1}^{m} x_n x_n + k\right] \quad \text{and} \quad c_k = \lim_{m\to\infty} c_k^{(m)}$$

(a) Show that $c_k^{(m)} = \left(1 - \frac{1}{N}\right)^2 (1-c)^k \left[\frac{1-(1-c)^m}{c}\right]$.

(b) Compute $\lim_{m\to\infty} c_k^{(m)}$.

7. Show that $EV\left(\sum_{n=1}^{\infty} n x_n\right) = \frac{u_1}{c}$.

8. Show that $EV\left(\sum_{n=1}^{\infty} \frac{x_n}{n} = \frac{1-\frac{1}{N}}{1-c} \log \frac{1}{c}\right)$.

9. **(a)** Prove that the formula for $P(T = k)$ in Section III is correct.

(b) Show that $EV(T) = u_1$.

(c) Show that $\mathrm{Var}(T) = u_1 + (1-2c)(u_1)^2$.

10. Define u_j by $u_j = \sum_{n=1}^{\infty} x_n x_{n+1} x_{n+2} \ldots x_{n+j-1}$ for $j = 1, 2, \ldots$.

Under what conditions is $x_n x_{n+1} x_{n+2} \ldots x_{n+j-1}$ zero?

11. For the sequence of labeled responses 11111001100001101000 . . . (all the rest zeros), show that

(a) $u_1 = 10$

(b) $u_2 = 6$

(c) $u_3 = 3$

(d) $u_4 = 2$

(e) $u_5 = 1$

(f) $r_1 = 1$

(g) $r_2 = 2$

(h) $r_3 = r_4 = 0$

(i) $r_5 = 1$

(j) $R = 4$

12. Show that the following relations hold true for the example of Exercise 11:

(a) $r_j = u_j - 2u_{j+1} + u_{j+2}$

(b) $R = u_1 - u_2$

13. Prove that the relations of Exercise 12 hold true for all sequences of labeled responses.

14. **(a)** Prove that

$$\Pr(x_{n+i} = 1 | x_n = 1, x_{n+1} = 1, \ldots, x_{n+i-1} = 1) = \Pr(x_{n+i} = 1 | x_{n+i-1} = 1)$$

(b) Show that $\Pr(x_{n+1} = 1 | x_n = 1) = (1-c)[1 - \frac{1}{N}]$.

15. Let $a = (1-c)(1-\frac{1}{N})$. Show that

(a) $EV(u_j) = u_1 a^{j-1}$

(b) $EV(R) = u_1(1-a)$

(c) $EV(r_j) = \mathrm{R}(1-a)a^{j-1}$

16. Let A_n be the random variable $A_n = (1 - x_{n+1})x_{n+1} + x_n(1 - x_{n+1})$.

(a) Show that $A_n = 1$, if $x_n = 1$ and $x_{n+1} = 0$.

(b) Show that $A_n = 1$ if $x_n = 0$ and $x_{n+1} = 1$.

(c) Show that $A_n = 0$ if $x_n = x_{n+1}$.

17. Let $A = \sum_{n=1}^{\infty} A_n$, where A_n is as defined in Exercise 16.

(a) Show that A counts the number of alternations of successes and failures.

(b) Evaluate the expected value of A, and show that

$$EV(A) = u_1\left[c + \frac{2(1-c)}{N}\right]$$

(c) What is $EV(A)$ if $N = 2$? Does the result make any intuitive sense to you?

18. Show that g, the probability that the first success occurs by guessing, is given by $g = \frac{1}{N} + \left(1 - \frac{1}{N}\right)(1 - c)\frac{1}{N} + \left(1 - \frac{1}{N}\right)^2(1 - c)^2\frac{1}{N} + \cdots + = \frac{1}{N(1-a)}$.

19. Let J be the random variable that is the number of errors before the first success. Show that

$$\Pr(J = i) = \begin{cases} \dfrac{1}{N} & \text{for } i = 0 \\ [1 - (1/N)](1 - \alpha)\alpha^{i-1} & \text{for } i \geq 1 \end{cases}$$

20. Show that the quantity b in the formula for $\Pr(\mathrm{T} = k)$ is the probability that no errors occur following a correct guess.

21. If p_1 denotes the probability that no error follows the first correct response, show that $p_1 = 1 - (1 - b)/N$.

22. Let W be the random variable that is the number of the trial on which the last error occurs.

(a) Show that

$$\Pr(W = k) = \begin{cases} \dfrac{b}{N} & \text{for } k = 0 \\ b[1 - (1/N)](1 - c)^{k-1} & \text{for } k \geq 1 \end{cases}$$

(b) Show that $EV(W) = \frac{bu_1}{c}$.

23. Some authors define the learning curve to be the graph of the proportion of *correct* responses as a function of the number of trials. Using this definition, sketch the learning curve predicted by the Bower model, using $c = .345$.

24. Does the design of Bower's experiment answer all the objections about the simplicity of the axioms? Which objections do you believe are most significant?

25. Compute $\mathrm{EV}(x_n x_{n+k} x_{n+2k})$.

26. Compute $\Pr(x_n = 0 | x_{n+k} = 1)$.

27. Find the probability that the subject is in state C by trial $n + k$ given that the last error occurred on trial n.

SUGGESTED PROJECTS

1. In the text and exercises formulas for some—but not all—of the predictions (A)–(N) of the Bower model are given. Discover and prove formulas for the remaining predictions.

2. Formulate and analyze a mathematical model for the following paired-associate learning situation. The experiment will be exactly the same as Bower's except that whenever a subject gives a response, the experimenter tells her only if the response is correct or incorrect. He does not tell her what the correct response is if she gives an incorrect one. Thus, there is much less frequent reinforcement of the connection between a stimulus element and its correct response. Which axioms of Bower's model would you retain? Which ones need modification? Compare the predictions of your model with those of Bower's.

3. The "linear model" for PAL is briefly mentioned in Section III. Find out what the assumptions and conclusion of this model are. Which predictions agree with those of the Bower model? In what learning situations would it be a more relevant model? Begin with the paper of Robert Bush and Saul Sternberg [1959] (see References) or the early chapters of Frank Restle and James G. Greeno [1970].

4. At the beginning this chapter, we divided the paired-associate learning task into two steps. Bower's model is concerned with the second part of the learning problem. Formulate a model for the learning process corresponding to the first step. Do Markov chains seem an appropriate modeling tool for this problem?

You can find a listing of references and suggestions for additional reading on the book's website, www.wiley.com/college/olinick

Epidemics

CHAPTER 14

Swords and lances, arrows, machine guns, and even high explosives have had far less power over the fate of nations than the typhus louse, the plague flea, and the yellow-fever mosquito.

—Hans Zinsser

I. Introduction

A. Epidemics and History

The period of Greek history from the end of the Persian Wars to the death of Alexander the Great (roughly 480 to 325 B.C.) was critical to the development of Western civilization. The creative achievements of the Greeks of this time in art, literature, philosophy, science, mathematics, and political science exerted an influence on Western cultural history unequaled by any other people.

Foremost among the Greek communities of 2,500 years ago was the city-state of Athens. No other state rivaled the extent of Athens's empire or its wealth, power, and intellectual and cultural activity, and none possessed so pure a democracy. The "Golden Age" of Athens coincided closely with the reign of Pericles, the most dominating personality of his time, who rose to power in 469 B.C. while still in his early thirties. The Golden Age began to tarnish, however, with the outbreak of the Peloponnesian War in 431 B.C. The war, which lasted a quarter-century, was essentially a series of military struggles between Athens and Sparta, the other predominant city-state. Athens was primarily a sea power with a strong navy, but it had a weak army compared to the Spartans, who had a strong army but no major fleet of ships. Pericles's strategy was to withdraw the population of the surrounding area into Athens, thus making his state invulnerable to land attack, and then to raid the coasts of his enemy.

The strategy worked well during the first year of the war, and the defeat of Sparta seemed inevitable. But the Athenian plan of bringing large numbers of people into the fortified city area had disastrous and unforeseen consequences. Overcrowding and unsanitary conditions provided an ideal setting for the spread of disease. In 430 B.C., an epidemic devastated Athens. The disease, whose exact nature is still unknown, was virulent and highly contagious. Between 30% and 60% of the Athenians died within six to eight days of contracting the illness. Many of those who survived the high fevers, violent coughs, and distressing vomiting of the disease fell victim to other medical complications.

Even those who lived on were often scarred by the epidemic and left with blindness, deformed arms or legs, and amnesia.

The contemporary Greek historian Thucydides, whose *History of the Peloponnesian War* is a classic of Western literature, provides a vivid account of the illness and its aftereffects: "Physicians, in ignorance of the nature of the disease, sought to apply remedies, but it was in vain, and they themselves were among the first victims, because they often came into contact with it. No human art was of any avail, and as to supplications in temples, inquiries of oracles and the like, they were utterly useless, and at last men were overpowered by the calamity and gave them all up. . . . The general character of the malady no words can describe, and the fury with which it fastened upon each sufferer was too much for human nature to endure." Many of the dead were left unburied, and birds and animals that preyed on the corpses became infected and spread the disease even further. Many Athenians committed suicide to escape the pain and suffering of the infection.

When it appeared that the epidemic had at last ended, Pericles sent his fleet to capture the Spartan-held stronghold at Potidaea. The ships had barely reached the sea when the plague broke out among the crews with such ferocity that they were forced to return to Athens. There were fresh outbreaks in 429 and 428 B.C., and Pericles himself fell victim. After his death, Athens never again found a leader of his stature and wisdom.

The plague of Athens was instrumental to the disintegration of the Athenian empire. The destruction of the fighting power of the navy and the disastrous reduction in population at home prevented Athens from achieving a swift victory over Sparta. The war dragged on for years, bringing eventual defeat for the Athenians.

Perhaps worse than the loss of life was the demoralization of the city-state that the plague brought in its wake. The descent of a highly civilized state into the depths of cruelty and desperation is one of the major themes of Thucydides's history. He records [1942] the lawlessness that swept through Athens:

> *Men who had hitherto concealed what they took pleasure in, now grew bolder. For seeing the sudden change—how the rich died in a moment, and those who had nothing immediately inherited their property—they reflected that life and riches were alike transitory, and they resolved to enjoy themselves while they could, and to think only of pleasure. Who would be willing to sacrifice himself to the law of honor when he knew not whether he would ever live to be held in honor? The pleasure of the moment and any sort of thing which conduced to it took the place both of honor and of expediency. No fear of Gods or law of man deterred a criminal. Those who saw all perishing alike thought that the worship or neglect of the Gods made no difference. For offences against human law no punishment was to be feared; no one would live long enough to be called to account. Already a far heavier sentence had been passed and was hanging over a man's head; before that fell, why should he not take a little pleasure?*

The Athenian plague is perhaps the earliest for which we have a detailed account of the influence of epidemics upon historical events. It is, however, but one of many disastrous situations. In the 14th century, an estimated 25 million deaths, in a population of 100 million Europeans, were attributed to an epidemic of bubonic plague. In 1520 the Aztecs suffered an epidemic of smallpox that resulted in the death of half their population of 3.5 million. When measles first came to the Fiji Islands in 1875 as a result of a trip to

Australia by the King of Fiji and his son, it caused the death of 40,000 people in a population of 150,000. In the three-year period from 1918 to 1921, there were an estimated 25 million cases of typhus in the Soviet Union and about one in ten victims died from the disease. In a worldwide epidemic of influenza in 1919, more than 20 million persons perished from the illness and subsequent attacks of pneumonia. The possibility that a similar strain of influenza virus might attack the United States in 1976 led the government to plan for the vaccination of the entire nation of more than 200 million people.

In the late 20th century, the AIDS epidemic caused worldwide alarm. AIDS (**acquired immune deficiency syndrome**) is a collection of symptoms and infections resulting from damage to the immune system caused by the human immunodeficiency virus (HIV). Late stages of the condition leave individuals prone to opportunistic infections and tumors. One of the most destructive epidemics in recorded history, AIDS has killed more 25 million people since 1981. In 2005 alone, AIDS claimed an estimated 2.4−3.3 million lives, of which more than 570,000 were children. Experts predict that the number of people with HIV will rise to 60 million by 2015.

Although treatments for AIDS and HIV exist to slow the progression of the virus, there is no known cure. Close to 40 million people are currently afflicted with AIDS. The disease has been particularly severe in sub-Saharan Africa, accounting for 70 percent of all AIDS deaths in 2011. In 2007, the U.S. president's adviser on AIDS, Dr. Anthony Fauci, reported that "[f]or every one person that you put in therapy, six new people get infected. So we're losing that game, the numbers game."

Between April 2009 and May 2010, the H1N1 influenza virus triggered a global epidemic of "swine flu." World Health Organization officials estimated the death toll at 284,500; other researchers concluded the number who died might have been more than half a million, as many victims without access to health facilities went uncounted. In summer 2013, health officials raised concerns about a possibly emerging new pandemic labeled Middle East Respiratory Syndrome (MERS). The MERS virus is a coronavirus, the same types that caused SARS. What is alarming is the fatality rate from MERS for the first 50 confirmed cases. Thirty of these patients, 60%, have died. The SARS epidemic of 2002−03 had a fatality rate of 10%; an avian influenza virus originating in China had a 25% rate. MERS apparently also has an incubation period of 9 to 12 days, leaving an extended period during which people can spread the disease without realizing that they are sick.

These examples, among hundreds of other similar ones, indicate clearly that epidemics are major public health problems requiring careful study and prompt action to protect the citizenry. In this chapter, we will present some simple mathematical models of the spread of infectious diseases. These models will serve as an introduction to the rapidly growing field of mathematical epidemiology in which mathematicians and biologists are working together to gain a better understanding of the spread and control of epidemics. "The real stimulus," writes Norman Bailey [1975], an internationally renowned leader in this field, "comes from the need to be able to influence in a rational way public health decision-making on the control of serious diseases that affect many hundreds of millions of people in the world today. When mathematical modeling is directed towards theoretical problems, which if solved would have practical implications for the control or eradication of disease, then it can be both intellectually satisfying and socially valuable."

In discussing his work on mathematical models of epidemics, the physicist Ronald Mickens addressed one of the advantages of models:

> *There are reasons why you want mathematical models. One is that if you have a good mathematical model, you can do things to the model that you can't do to a human being. For example, you might want to investigate various strategies for giving a very effective measles vaccine. . . . But we know that measles has not been eliminated, and one reason is that, you have to vaccinate effectively ninety-five percent of all the susceptibles. . . . It's impossible to vaccinate ninety-five percent. . . . Well, there are parents who don't want their kids vaccinated because of the very, very, very low probability that something bad may happen to them. There are others who won't do it because of religious reasons. . . . But the mere existence of a strategy doesn't mean that you can carry it out. It may be unethical, at least for our society, and so one of the things you do with these models is to look at various kinds of strategies and then you can hand this off to somebody in public policy, and they can decide well, we can't do that because the society would not allow this to happen.*

In this chapter, we will present some of the features of infectious diseases that ought to be incorporated into realistic mathematical models, develop several deterministic models and a stochastic model and discuss their relationships, and conclude with a brief sketch of the development of mathematical epidemiology. We also present variations of the classic models for the spread of infectious diseases that are being used to model the dissemination of rumors, the persistence of urban legends, and the dynamics of such problems as problem drinking, spousal abuse, and eating disorders.

B. Some Features of Epidemics

The spread of an infectious disease among a population can be a complicated process with many possible variations. Consider first a single individual who may be infected by some contagious pathogenic agent. The organism may enter his body through the bite of a flea (as in bubonic plague) or mosquito (yellow fever), through intimate personal contact with another infected person (HIV), by airborne agents spread by coughing or sneezing (pneumonia), or by drinking contaminated water (typhoid fever).

After initial infection, there may be a *latent period* during which the individual exhibits no symptoms of the disease and cannot transmit it to others. The latent period is followed by an *infectious period* when he can pass on the illness. These two periods may be overlapped by an *incubation period:* the time between initial infection and first appearance of physical symptoms. Thus, an individual may be transmitting a disease to others during a period when he and others are unaware that he is sick. This is characteristic of some diseases, such as chickenpox and measles, now common mainly among younger children.

Once the symptoms appear, the affected individual may continue to be an active transmitter, especially if the disease is a mild one such as a cold or minor respiratory infection. On the other hand, the individual may be withdrawn from the general population temporarily (by quarantine or hospitalization, for example) or permanently (through death). The chances of recovery or death vary from day to day during the various stages of the illness, as do the chances that the individual will convey the infection to previously unaffected people.

If the individual recovers from the disease, there are still many possible scenarios. A permanent *immunity* to the disease may be acquired so that there is never again susceptibility to the symptoms even if there is reinfection. The immunity may be of such a nature that the individual can no longer even transmit the disease to others, or he may become a *carrier*: a person who can spread the illness but who is otherwise unaffected by it. The immunity may be temporary so that there is no susceptibility to the disease again for many months or years (tetanus), or there may be no immunity at all: the so-called "English sweating sickness" that ravaged Western Europe in the late 15th and early 16th centuries attacked some individuals two or more times in brief succession after they had recovered from an initial bout of the disease's associated tremors, fever, cardiac pain, vomiting, severe headache, and stupor.

As an epidemic spreads through a local community, city, nation, or continent, the number of unaffected members becomes reduced. In due course of time, the epidemic may appear to end, as no new cases of the disease are observed. In an early paper in the history of mathematical models of epidemics, two Edinburgh researchers, William O. Kermack and Anderson G. McKendrick [1927], posed the fundamental goal of such models. "One of the most important problems in epidemiology," they wrote, "is to ascertain whether this termination occurs only when no susceptible individuals are left, or whether the interplay of the various factors on infectivity, recovery and mortality, may result in termination, whilst many susceptible individuals are still present in the unaffected population."

If it can be shown that a particular epidemic will end when only a small proportion of the potentially susceptible members of the community have been affected, then there may be little cause for panic or widespread emergency public-health measures. Conversely, it is important to know early in the growth of an epidemic of a disease with a high mortality rate that large numbers of the population may become victims.

Even when a number of simplifying assumptions are made, mathematical models of epidemics tend to be quite complex and require advanced analytic and probabilistic techniques to solve. Even some of the simpler models give rise to mathematical problems that have yet to be solved. While waiting for the mathematicians to solve these problems, modelers must resort to approximate results or computer simulations (see Chapter 15) to derive their predictions. Because of the technical mathematical difficulties posed by many models, this chapter will concentrate only on some very simple models of epidemics. Even though the models are simple, they do yield qualitative results that are consistent with observations and that are helpful to biologists and public health officials.

II. Deterministic Models

A. Basic Assumptions

In studying a community subject to a possible epidemic, it is convenient to partition the population into four mutually exclusive subgroups:

1. The *susceptibles* (S), those persons who are currently uninfected, but may become infected
2. The *latently infected* (L), those who are currently infected, but not yet capable of transmitting the disease to others

3. The *infectives* (I), those who are currently infected and capable of spreading the infection
4. The *removeds* (R), those persons who have had the disease and are dead, or who have recovered and are permanently immune, or who are isolated until death, recovery, or permanent immunity occur

The numbers of persons, S, L, I, and R, in each category change with time. We will study mathematical models that attempt to discover how these numbers fluctuate with respect to time, denoted as usual by t, and with respect to each other.

Since the course of most epidemics is usually short (a few weeks or months) compared to the normal life span of an individual, a reasonable simplifying assumption is that the population of the community remains constant—except, of course, as it is lowered by deaths due to the epidemic disease itself. Suppose, then, that there are no births, no deaths from other causes, and no immigration or emigration during the course of the epidemic. This initial assumption is stated mathematically as the following axiom:

AXIOM 1 There is positive constant N such that $S(t) + L(t) + I(t) + R(t) = N$ for all t. For the simplification of some formulas, this equation is frequently written as

$$S + L + I + R = N \tag{1}$$

The second assumption is also one that is basic to almost all mathematical models of epidemics: *the rate of change of the susceptible population is proportional to the rate of contact between susceptibles and infectives*. If the amount of human interaction is great, the epidemic spreads more quickly. History records the rapid spread of many diseases in crowded cities or army or refugee camps, while showing that epidemics move more slowly through isolated rural areas with lower population densities. For simple models, it is usually assumed that the rate of contact is directly proportional to the population of susceptibles and infectives. In mathematical terms, the second basic assumption is the following:

AXIOM 2 There is a positive constant β such that

$$dS/dt = S'(t) = -\beta I(t)S(t) \quad \text{for all } t. \tag{2}$$

The constant β is called the *infection rate*. Note that $S'(t)$ is always negative (or possibly zero), since the number of persons who have not yet caught the disease, (S), can only decrease with time.

There is one final assumption that is common to the models we present in this chapter. We suppose that the disease being investigated has a latency period that is negligibly short—that is, an individual can transmit the disease essentially as soon as he is infected. Typhus fever presents such a possibility. Typhus is often spread from person to person through the bite of a hair louse. A louse carrying the disease may leave the body of a person it has just bitten and move on to a nearby person at any moment. In the language of the variables of the deterministic models, we have this axiom:

AXIOM 3 $L(t) = 0$ for all t.

B. A Simple Epidemic Model

This section investigates a deterministic model of a simple epidemic. In addition to Axioms 1−3, we make one further simplification. We assume that there is no removal from the population; the population of removeds remains at 0. In a human population, such an assumption might be justified for an illness such as a mild cold epidemic in a college dormitory. None of the affected students dies, acquires permanent immunity, or is sick enough to be isolated. The assumption is commonly valid also in many cases of disease in animal or plant population, where dead or diseased members in a natural environment are not removed. Formally stated, the assumption is

AXIOM 4 $R(t) = 0$ for all t.

Using the tools of elementary calculus, we can easily analyze a mathematical model for an epidemic satisfying Axioms 1−4. This model is sometimes called an *SI* model since it involves susceptibles and infectives only. Axioms 1, 3, and 4 give the basic relation between susceptibles and infectives:

$$S + I = N \tag{3}$$

For convenience, assume the epidemic starts at time $t = 0$ with a single infected person, so the initial conditions are

$$I(0) = I_0 = 1 \quad \text{and} \quad S(0) = S_0 = N - 1 \tag{4}$$

The analysis of this simple epidemic begins by using Eq. (3) to rewrite Axiom 2 as the differential equation

$$\frac{dI}{dt} = \frac{d(N - S)}{dt} = -\frac{dS}{dt} = \beta IS = \beta I(N - I) \tag{5}$$

so that the number of infectives is governed by the differential equation

$$\frac{dI}{dt} = \beta IS = \beta I(N - I), I(0) = 1 \tag{6}$$

Eq. (6) is a differential equation for logistic growth. We studied such equations extensively in Chapter 3. They are solved by separating the variables (I and t in this case) and integrating using a partial fraction decomposition. Recall that we are using log for the natural logarithm.

$$\int \frac{dI}{I(N - I)} dt = \int \beta dt$$

or

$$\int \left(\frac{1}{I} + \frac{1}{N - I} \right) dt = \int \beta N dt$$

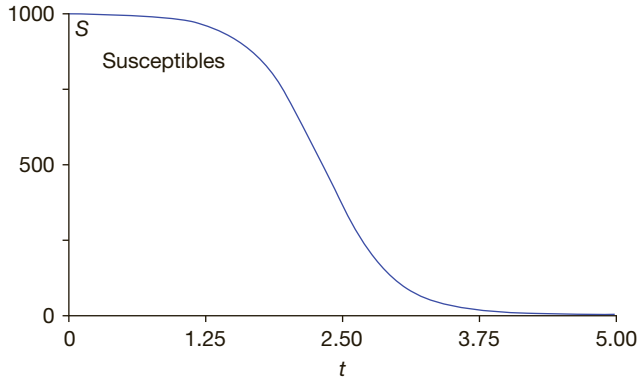


FIGURE 14.1 The graph of $S(t)$ as a function of t for the simple deterministic model, with $N = 1001$ and $\beta = .003$.

so that

$$\log I - \log(N - I) = \beta N t + \text{ constant}$$

which may be written in the equivalent form

$$\frac{I}{N - I} = Ke^{\beta N t} \tag{7}$$

where the constant K is found, using Eq. (4), to be $\frac{1}{N-1}$. Eq. (7) may then be rewritten as

$$I(t) = \frac{N}{1 + (N - 1)e^{-\beta N t}} \tag{8}$$

Since β is positive, it is apparent from Eq. (8) that

$$\lim_{t \to \infty} I(t) = N \tag{9}$$

Thus, the model predicts that everyone in the population will eventually contract the disease. Since $S + I = N$, we have $S = N - I$, or

$$S(t) = N - I(t) = \frac{N(N - 1)}{e^{\beta N t} + (N - 1)} = \frac{N}{1 + \frac{e^{\beta N t}}{N - 1}} \tag{10}$$

The graphs of S and I as functions of t are given in Figs. 14.1 and 14.2 for the case $N = 1001$ and $\beta = .003$. According to this model, if a single infective person enters a community of 1,000 susceptibles, then at the end of four time units, only about six healthy people will be left. The decline in the number of susceptibles is also represented in Table 14.1.

The data collected in an epidemic often consists in the number of **new** cases of the disease reported each day or each week. The *rate* at which new cases arise is the

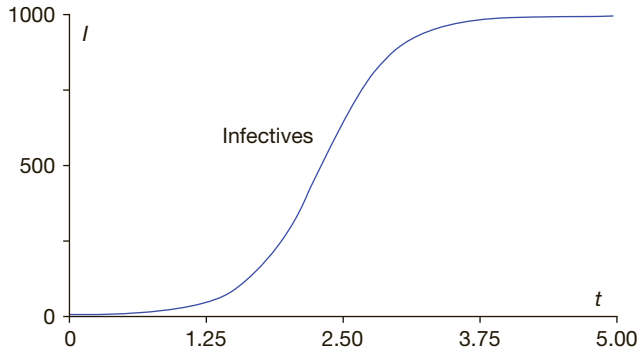


FIGURE 14.2 The graph of $I(t)$ as a function of t for the simple deterministic model, with $N = 1001$ and $\beta = .003$.

Table 14.1 Decline in the number of susceptibles for selected times in the simple epidemic with $N = 1001$ and $\beta = .003$. The number $S(t)$ is computed from Eq. (10).

Time t	Susceptibles $S(t)$
0	1,000.0
.5	996.5
1	981.2
1.5	918.0
2	712.0
2.5	354.8
3	109.1
3.5	26.6
4	6.0
4.5	1.4

derivative of I with respect to time t. From Eq. (8), this rate can be computed explicitly as a function of t:

$$I'(t) = \frac{dI}{dt} = \frac{N^2(N-1)\beta e^{-\beta N t}}{(1+(N-1)e^{-\beta N t})^2} \tag{11}$$

The graph of $I'(t)$ as a function of t is called the *epidemic curve*. The epidemic curve for $N = 1001$ and $\beta = .003$ is shown in Fig. 14.3. At the start of the epidemic $(t = 0)$, the derivative has value

$$I'(0) = \frac{N^2(N-1)\beta}{(1+(N-1))^2} = \beta(N-1) \tag{12}$$

Assuming that β has been computed using time measured in weekly units, the particular epidemic of Fig. 14.3 would begin with about three cases per week. Table 14.2 shows additional numerical data on the values of $I'(t)$.

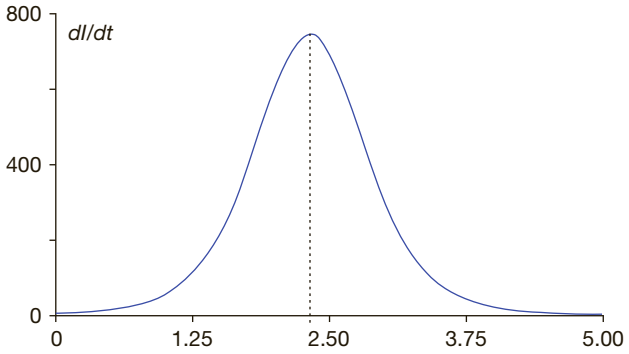


FIGURE 14.3 The epidemic curve for the simple deterministic model, with $N = 1001$ and $\beta = .003$.

Table 14.2 Fluctuations in the rate of new cases as measured by $I'(t)$, using Eq. (11) with $N = 1001$ and $\beta = .003$. The values of $I'(t)$ have been rounded to the nearest integer.

Time t	Rate of New Cases, $I'(t)$
0	3
0.5	13
1	58
1.5	229
2	617
2.5	688
3	292
3.5	78
4	18
4.5	4

The maximum value for the rate of new cases occurs at the maximum height of the epidemic curve. The time at which this maximum occurs represents a critical moment in the history of the epidemic: it is spreading most rapidly at this instant, and the consequent demand for medical services and personnel may be at its peak. To find the maximum value of $I'(t)$, consider its derivative

$$I''(t) = \left[\frac{(N-1)N^3\beta^2 e^{-\beta Nt}}{(1+(N-1)e^{-\beta Nt})^3}\right]\left[(N-1)e^{-\beta Nt} - 1\right] \tag{13}$$

Now the factor inside the first square brackets in Eq. (13) remains positive for all values of t. Thus, the sign of the second derivative $I''(t)$ depends on the sign of the remaining factor

$$\frac{N-1}{e^{\beta Nt}} - 1 \tag{14}$$

This factor is positive when $t = 0$ and tends toward -1 as t gets large. (Why?) This implies that the epidemic curve reaches its maximum height when this factor is zero. This occurs when

$$e^{\beta N t} = (N - 1) \tag{15}$$

that is

$$t_{max} = \frac{\log(N-1)}{\beta N} \tag{16}$$

At this moment, the rate of new infections is

$$I'(t_{max}) = \frac{N^2\beta}{4} \tag{17}$$

and the number of infected individuals is

$$I(t_{max}) = \frac{N}{2} \tag{18}$$

The graph of the epidemic curve of Fig. 14.3 appears to be symmetric about the vertical line through t_{max} (with $N = 1001$ and $\beta = .003$, the value of t_{max} is about 2.3). The apparent symmetry is real and holds in general for the simple epidemic model; see Exercise 6.

Note finally from Eq. (16) that the smaller the value of β is, the longer it takes the epidemic curve to reach its peak. Thus, the model shows that the more densely crowded a population is, the faster will an epidemic spread through the community.

Although this simple model has a number of interesting results consistent with real-world observations, it makes one prediction that rarely is correct in actual epidemics. The model asserts that before the epidemic runs its course everyone will contract the illness. In the real world, the epidemic ends—in the sense that no new infectives are seen—while there are still many susceptibles in the population. In the next section, we will turn to a model that is consistent with this observation, but first we will examine a discrete version of the simpler model.

C. A Discrete Version of the Simple Epidemic Model

Real-world data on epidemics is collected at discrete time intervals, usually every day, week, or month depending on the nature and severity of the disease. It's useful then to examine the discrete dynamical version of the simple epidemic model.

If S_k, I_k, L_k, R_k denote the number of susceptibles, infectives, latents, and removeds, respectively, at the start of the kth time interval, then the axioms of the continuous simple model become, in the discrete case:

DISCRETE AXIOM 1: There is a positive constant N such that $S_k + L_k + I_k + R_k = N$ for all $k \geq 0$.

DISCRETE AXIOM 2: There is a positive constant β such that $S_{k+1} - S_k = -\beta I_k S_k$ for all $k \geq 0$.

DISCRETE AXIOM 3: $L_k = 0$ for all $k \geq 0$.

DISCRETE AXIOM 4: $R_k = 0$ for all $k \geq 0$.

As a consequence of these axioms, we have $S_k + I_k = N$ so that $S_k = N - I_k$ and thus,

$$I_{k+1} - I_k = (N - S_{k+1}) - (N - S_k) = S_k - S_{k+1} = \beta I_k S_k = \beta I_k (N - I_k)$$

and hence,

$$I_{k+1} = I_k + \beta I_k (N - I_k).$$

We need to make one important modification in this last equation, since the right-hand side could exceed N for some value of k, but the population of infectives can never be larger than the total population. Thus, we modify our final relationship to get the central equation of our discrete model:

$$\boldsymbol{I_{k+1} = \text{minimum}(I_k + \beta I_k (N - I_k), N)}$$
$$\boldsymbol{I_0 = \text{initial number of infectives}}$$

It turns out that there are two cases to consider, depending on the size of β. If $\beta \leq 1/N$, then some positive fraction of susceptibles remains in the population at every period. The infective population will approach N from below in the limit. In this case, $I_{k+1} = I_k + \beta I_k (N - \mathrm{I}_k)$ for all k.

If $\beta > 1/N$, then after a finite number of time periods, everyone is infected. At some value of k, we will have minimum $(I_k + \beta I_k (N - I_k), N) = N$.

To prove these claims, consider the function $f(x) = x + \beta x(N - x)$, the graph of which is a parabola, opening downward. (See Figs. 14.4 or 14.5.) Note that

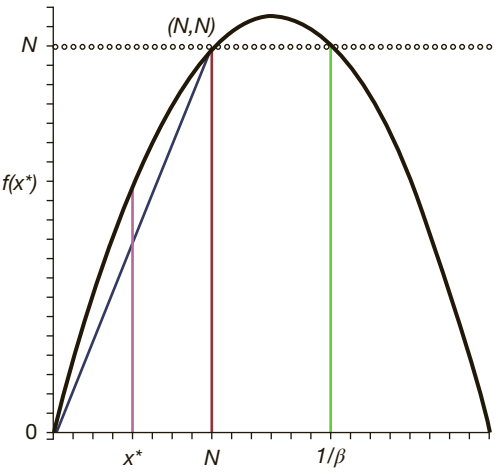


FIGURE 14.4 The graph of $f(x) = x + \beta x(N - x)$, in the case $\beta \leq 1/N$.

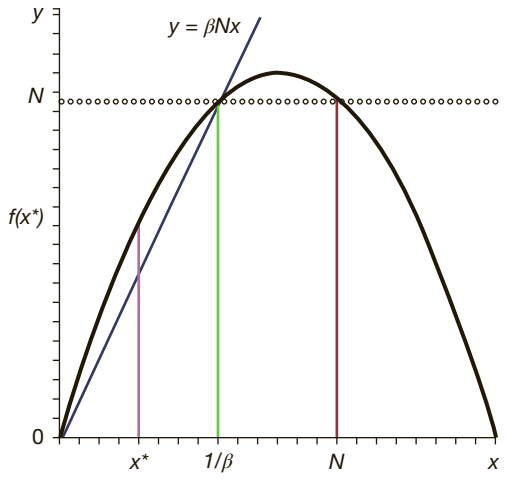


FIGURE 14.5 The graph of $f(x) = x + \beta x(N - x)$, in the case $\beta \leq 1/N$.

$f(N) = N + \beta N(N - N) = N$ and that $f(1/\beta) = 1/\beta + \beta(1/\beta)(N - 1/\beta) = 1/\beta + N - 1/\beta = N$. The maximum value of f occurs at the average of $1/\beta$ and N.

Observe also that $f(I_k) = I_k + \beta I_k(N - I_k)$.

Case 1: $\beta \leq 1/N$

In this case, $N\beta \leq 1$, or, equivalently, $N \leq 1/\beta$.

Consider any x^* with $0 < x^* < N$. Because f is increasing on $[0, N]$, we have $f(x^*) < f(N) = N$. Since the graph of f is concave down, we also have $f(x^*) > x^*$. Thus, for any x^* in $[0, N]$, we have $x^* < f(x^*) < N$.

If $0 < I_0 < N$, then $I_{k+1} = \text{minimum}(I_k + \beta I_k(N - I_k), N) = \text{minimum}(f(I_k), N) = f(I_k)$ for all k. In summary, if $\beta N \leq 1$, then $I_k + \beta I_k(N - I_k)$ is always less than N, so I_{k+1} always remains below N. Some positive fraction of susceptibles remains in the population at every period.

Case 2: $\beta > 1/N$

In this case, $\beta N > 1$, or, equivalently, $N > 1/\beta$. See Fig. 14.5.

Since f is concave down, the line joining two points on the graph of f lies below the graph of the curve. Since the line between $(0, f(0)) = (0, 0)$ and $(1/\beta, f(1/\beta)) = (1/\beta, N)$ has equation $y = \beta N x$, we have $f(x^*) > \beta N x^*$ for any x^* between 0 and $1/\beta$. Moreover, since f is increasing on $[0,1/\beta]$, we will also have $f(x^*) < f(1/\beta) = N$ for such an x^*.

Hence, if $0 < x_0 < 1/\beta$, then $x_1 = f(x_0) > \beta N x_0$ and $x_2 = f(x_1) > \beta N x_1 > (\beta N)^2 x_0$, $x_3 = f(x_2) > \beta N x_2 > (\beta N)^3 x_0$, and, in general, $x_t = f(x_{t-1}) > (\beta N)^t x_0$.

Since $\beta N > 1$, the powers $(\beta N)^t$ grow arbitrarily large as t increases. Thus, there will be a smallest positive integer T such that

$$(\beta N)^{\mathrm{T}} x_0 < 1/\beta < (\beta N)^{T+1} x_0 < x_{T+1}$$

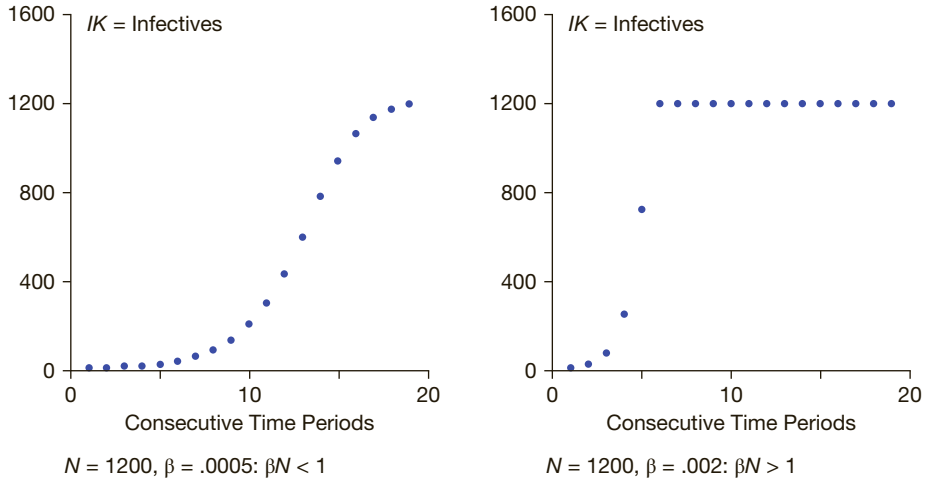


FIGURE 14.6 Two discrete time versions of the simple epidemic model.

With $I_0 = x_0$, we would then have

$$I_{T+2} = \text{minimum}(I_{T+1} + \beta I_{T+1}(N - I_{T+1}), N) = N$$

and hence $I_k = N$ for all $k \geq T + 2$.

In summary, if $\beta N > 1$—that is, β is sufficiently large—then after a finite number of time periods, everyone is infected.

In Fig. 14.6, we show the results of two typical runs of the discrete version of the simple epidemic model. In the first, we picked a small value of β so that $\beta N < 1$. In the second, we made β sufficiently large that $\beta N > 1$.

D. A More General Epidemic Model

Description of the SIR Model

The simple epidemic model just studied assumes that once an individual becomes an infective, he remains one from then on. Thus, the population of infectives can only increase. In this section, we examine a model in which the subgroup of infectives is *increased* by the introduction of formerly susceptible persons and is *decreased* by the removal of some individuals who either die from the disease, recover from it and acquire permanent immunity, or are isolated from the remaining population during the course of their illness. In place of Axiom 4, this more general model assumes that individuals are removed from the infective class at a rate that is proportional to the number of infectives—that is, there is a constant removal rate per person. In mathematical terms, the new axiom is as follows:

AXIOM 4* There is a positive constant r such that $R'(t) = rI(t)$ for all t.

The constant r is called the *removal rate*, and the ratio $p = r/\beta$ is the *relative removal rate*.

Epidemic researchers use the term **SIR model** for one that incorporates Axiom 4*. For a model satisfying Axioms 1, 2, 3, and 4*, we have

$$I = N - S - R \tag{19}$$

so that

$$\frac{dI}{dt} = 0 - \frac{dS}{dt} - \frac{dR}{dt} = \beta IS - rI \tag{20}$$

Assuming that the epidemic starts in a community of N persons with a positive number I_0 of infectives and with $S_0 = N - I_0$ susceptibles, the mathematical model is the system of differential equations

$$\frac{dS}{dt} = -\beta SI, \quad \beta > 0 \tag{21.1}$$

$$\frac{dI}{dt} = \beta SI - rI, \quad r > 0 \tag{21.2}$$

$$\frac{dR}{dt} = rI \tag{21.3}$$

with initial conditions

$$S(0) = S_0, \quad I(0) = I_0 > 0, \quad R(0) = R_0 = 0$$

and the relation

$$S(t) + I(t) + R(t) = N \text{ for all } t \geq 0$$

In the rest of this section, we shall explore many of the conclusions that can be derived from this model.

Qualitative Behavior of *R*, *S*, and *I*

Since r is positive and I is nonnegative, $dR/dt = rI$ is always nonnegative, so R is a monotonic nondecreasing function of t. In fact, R is a strictly increasing function except at the time the number of infectives drops to zero.

Since $\beta > 0$ and S and I are nonnegative, we have $dS/dt \leq 0$ for all t. Thus, the number of susceptibles is a monotonic nonincreasing function: the population of susceptibles can only decrease as time goes on.

Write the rate of change of infectives as $dI/dt = I(\beta S - r)$. The sign of this rate then depends on the sign of $(\beta S - r)$. The number of infectives can increase only at times when dI/dt is positive—that is, at times when

$$S > \frac{r}{\beta} = p \tag{22}$$

In particular, if the initial population level of susceptibles, S_0, is below the relative removal rate p, then there is no epidemic. The number of infectives is always less than the original number (I_0) and decreases as time goes on. The disease dies out as infected individuals are being removed (by recovery or death) at a faster rate than they are becoming sources of further infection. This is what epidemiologists term a *threshold phenomenon*. There is a critical value that the initial susceptible population must exceed for there to be an epidemic. For example, if a sufficiently high percentage of the population has been successfully vaccinated against the disease, then there will be no epidemic. Alternatively, holding the susceptible population fixed, infection can spread only if the relative removal rate is sufficiently small: an epidemic can be halted by increasing the relative removal rate p.

Limits of *R*, *S*, and *I*

The numbers of removeds, susceptibles, and infectives must always lie between 0 and N, the total size of the community. The function $R(t)$ is bounded above by N and is monotonically nondecreasing, so the number of removeds reaches a limit as time goes on—that is, $\lim_{t\to\infty} R(t)$ exists. Denote this number by R_∞. Thus,

$$\lim_{t\to\infty} R(t) = R_\infty \leq \mathrm{N} \tag{23}$$

Similarly $S(t)$ is a nonincreasing function of t that must remain greater than or equal to 0, so it has a limit also as t increases. There is a nonnegative number $S\infty$ such that

$$\lim_{t\to\infty} S(t) = S_\infty \geq 0 \tag{24}$$

Consider the limiting value for the number of infectives. From Eq. (19),

$$\lim_{t\to\infty} I(t) = \lim_{t\to\infty} N - S(t) - R(t) = N - \lim_{t\to\infty} S(t) - \lim_{t\to\infty} R(t) = N - S_\infty - R_\infty \tag{25}$$

Let I_∞ denote this limiting value.

Note that the number $\frac{R_\infty}{N}$ is the proportion of the population that eventually has the disease. It provides a convenient measure of the intensity of the epidemic.

Relation of *R* and *S*

The relationship between the number of susceptibles and the number of removeds during the course of the epidemic becomes more apparent if we use Eqs. (21.1) and (21.3) to write

$$\frac{dS}{dR} = -\frac{\beta SI}{rI} = -\frac{\beta}{r}S = -\frac{1}{p}S \tag{26}$$

which is a relation that holds whenever there are still infectives in the population. The differential equation (26) is easily solved to obtain

$$S = S_0 \mathrm{e}^{(-1/p)R} \tag{27}$$

Since $R(t) \le R_\infty \le N$ for all t, we have $e^{(-1/p)R} \ge e^{(-1/p)N}$ so that

$$S(t) \ge S_0 e^{(-1/p)N}, \quad \text{for all } t \tag{28}$$

Since the right-hand side of Eq. (28) is a strictly positive number, we have

$$S_\infty = \lim_{t\to\infty} S(t) > 0 \tag{29}$$

Here is a crucial prediction of this model: there will always be some people (S_∞ of them) in the community who escape the disease. The epidemic will die out, but not because there aren't susceptible individuals left.

Relation of *S* and *I*

Examine next the relationship between the number of susceptibles and the number of infectives during the epidemic. Eqs. (21.1) and (21.2) define an autonomous system of differential equations (see Chapter 4) for which the only critical points lie on the line $I = 0$. We are interested in orbits of the system that lie in the first quadrant of the (S,I)-plane. From the two equations, we have

$$\frac{dI}{dS} = \frac{I(\beta S - r)}{-\beta SI} = -1 + \frac{r}{\beta S} = -1 + \frac{p}{S} \tag{30}$$

whenever $I \neq 0$.

Separating the variables in the differential Eq. (30) and integrating yields

$$\int 1 dI = \int \left(-1 + \frac{p}{S}\right) dS \tag{31}$$

so that

$$I = -S + p\log S + C \tag{32}$$

where C is a constant. At time $t = 0$, there are I_0 infectives and S_0 susceptibles so that

$$C = I_0 + S_0 - p\log S_0 = N - p\log S_0 \tag{33}$$

This value of C gives

$$I = N - S + p\log\left(\frac{S}{S_0}\right) \tag{34}$$

Thus, the orbits for solutions of the autonomous system lie along the curve with equation $I = g(S) = N - S + p\log(S/S_0)$. Since $g(S_0) = I_0 > 0$ and $\lim_{S\to 0^+} g(S) = -\infty$, the curve crosses the line $I = 0$ at some positive value of S less than S_0. Since the only critical points for the system lie on the line $I = 0$, the orbit must approach $(S_\infty,0)$ as t increases. Thus, $I_\infty = 0$.

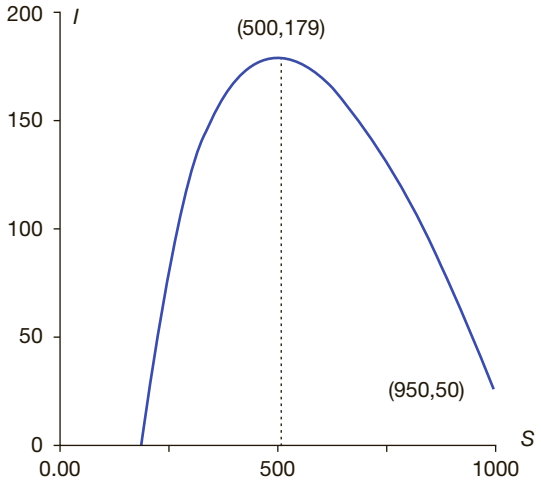


FIGURE 14.7 The graph of $I = N - S + p\log(S/S_0)$ with $N = 1000$, $p = 500$, $S_0 = 950$. Here S_∞ is about 186, and the maximum value of I is about 179.

The orbit is traced out from right to left as t increases since S is a nonincreasing function of time. From Eq. (30), we have, for $g(S) = N - S + p\log(S/S_0)$,

$$g''(S) = \frac{d^2I}{dS^2} = -\frac{p}{S^2} \tag{35}$$

which is always negative. Thus, the graph of I as a function of S is concave down and reaches its maximum when $dI/dS = 0$. From Eq. (30), this happens when $S = p$. The relation between I and S is shown in Fig. 14.7. The initial state of the population is represented by the point (S_0, I_0) on this curve. If this point falls to the left of the line $S = p$, then no epidemic occurs: $I(t)$ shrinks monotonically toward zero. On the other hand, if $S_0 > p$, then the number of infectives increases initially until S passes below p after which the number of infectives again falls toward zero.

Finding S

To locate S_∞, more precisely, note that Eq. (25) gives $S_\infty = N - R_\infty - I_\infty$, but we have just seen that $I_\infty = 0$, so that

$$S_\infty = N - R_\infty \tag{36}$$

The relation between S and R given by Eq. (27) is

$$S(t) = S_0 e^{(-1/p)R(t)} \tag{27}$$

and if we let $t \to \infty$ in Eq. (27), we have

$$S_\infty = S_0 e^{(-1/p)R_\infty} \tag{37}$$

Combining this relationship with Eq. (36) produces

$$S_\infty = S_0 e^{(-1/p)(N-S_\infty)} \tag{38}$$

so that S_∞ is a solution of the equation

$$S_0 e^{(-1/p)(N-x)} - x = 0 \tag{39}$$

Unfortunately, we cannot solve this equation analytically for x as an explicit function of S_0, p, and N, but we can show that there is a unique positive solution, and we can approximate its value to any desired accuracy. Toward these ends, define the function

$$f(x) = S_0 e^{(-1/p)(N-x)} - x \text{ for } x \geq 0.$$

Note that

$$f(0) = S_0 e^{(1/p)N} > 0$$

while

$$f(N) = S_0 - N < 0$$

Since f is a continuous function of x, the Intermediate Value Theorem of elementary calculus asserts that there is at least one number x^* between 0 and N for which $f(x^*) = 0$. Hence, there is at least one positive root of the equation.

Consider next the derivative of f:

$$f'(x) = \frac{S_0}{p} e^{(-1/p)(N-x)} - 1 = \frac{f(x) + x}{p} - 1$$

so that

$$f'(x^*) = \frac{f(x^*) + x^*}{p} - 1 = \frac{0 + x^*}{p} - 1 = \frac{x^*}{p} - 1$$

Since $x^* = S_\infty < p, f'(x^*) < 0$. If there are two or more roots, then Rolle's Theorem guarantees there is a point between the roots at which the derivative is 0. But the derivative is negative at both roots and the second derivative $f''(x) = S_0(1/p)^2 e^{(-1/p)(N-x)}$ is always positive, so the derivative must always be negative between the roots. This contradiction shows that there cannot be more than one root.

The preceding discussion establishes the fundamental theorem for this general epidemic model:

THEOREM (THRESHOLD THEOREM OF EPIDEMIOLOGY) If $S_0 < r/\beta$, then $I(t)$ goes monotonically to zero. If $S_0 > r/\beta$, then the number of infectives increases as t increases and then tends monotonically to zero. The limit of $S(t)$ as $t \to \infty$ exists and is the unique positive root of the equation

$$S_0 e^{(\frac{\beta}{r})(N-x)} - x = 0.$$

Approximation of S_∞

This last equation cannot be solved for x in closed form, but various methods are available to approximate its value. We discuss a particularly simple one, based on the Intermediate Value Theorem, here.

The Bisection Technique.

Suppose we have a continuous function defined on the closed interval $[0, N]$ with the property that $f(0) > 0$ and $f(N) < 0$. By the Intermediate Value Theorem, there is a root of the equation $f(x) = 0$ somewhere on this interval of length of N. Split this interval into two equal parts, subintervals $[0, N/2]$ and $[N/2, N]$. If $f(N/2) < 0$, then there is a root between 0 and $N/2$ while if $f(N/2) > 0$, there is a root between $N/2$ and N. In either case, we have narrowed the search for a root to an interval of length $N/2$. By examining the midpoint of this interval in the same manner, we can narrow the search down to an interval of length $N/2^2$. Continuing this process k times produces an interval of length $N/2^k$, which contains a root of $f(x) = 0$. By choosing k sufficiently large, we can find a numerical value for the root to a desired degree of accuracy.

As an example, consider a population of 1,001 individuals with a single infective at time 0. With $\beta = .001$ and $r = .9$, this "bisection" process requires 10 steps to obtain a value of S_∞ equal to 799, which is accurate to the nearest integer. In such an epidemic, about 80% of the population would not be affected by the disease.

Relation between R and t

Note that we have not derived explicit solutions of the differential equations of (21.1)–(21.3) in the form of functions of time. In this section, we will find an approximate solution for $R(t)$. We concentrate on the number of removeds, since it is frequently not possible to determine when an individual is first infected, but it is usually easier to observe when he has been removed.

Rewrite Eq. (21.3), $dR/dt = rI$ as

$$dR/dt = r(N - S - R) \tag{40}$$

which, by Eq. (27), can be represented as

$$dR/dt = r(N - R - S_0 e^{(-1/p)R}) \tag{41}$$

This is a single differential equation in the variables R and t, with initial condition $R(0) = 0$. Although an exact solution of this equation is possible, at least in parametric terms, the necessary work is rather complicated. We will illustrate an approximate approach that yields a number of interesting properties and is typical of the way some analytic models are studied.

A Taylor series approximation for the exponential function e^x is given by choosing an initial string of terms from the series

$$e^x = 1 + x + \frac{x^2}{2} + \frac{x^3}{3!} + \cdots + \frac{x^k}{k!} + \cdots$$

Approximating the function with the first three terms gives

$$e^x \sim 1 + x + \frac{x^2}{2}$$

and with $x = (-1/p)R$, this yields in place of Eq. (41),

$$\frac{dR}{dt} \sim r\left(N - R - S_0\left(1 - \frac{R}{p} + \frac{R^2}{2p^2}\right)\right) \tag{42}$$

Ultimately when the epidemic ends, $dR/dt = 0$. If the original infective population is very small, so that the number of initial susceptibles is close to the total population, we have $S_o \sim N$, so that $dR/dt = 0$ and $R = R_\infty$ and

$$r\left(S_0 - R_\infty - S_0 + \frac{S_0 R_\infty}{p} - \frac{S_0 R_\infty^2}{2p^2}\right) \sim 0$$

which occurs when

$$R_\infty \sim 2p\left(1 - \frac{p}{S_0}\right) \tag{43}$$

This expression gives an approximate measure of the total number of people who have contracted the disease. Since an epidemic occurs only if $S_0 > p$, let v be the positive number $S_0 - p$, so that $S_o = p + v$. Then Eq. (43) can be written

$$R_\infty \sim \frac{2pv}{p + v} \tag{44}$$

If v is small in comparison to p, then $p/(p + v)$ is nearly 1, and $R_\infty \sim 2v$. In other words, the total size of the epidemic is about $2v$ cases. The initial population of susceptibles, $p + v$, is thus reduced to $p + v - 2v = p - v$. The susceptible population is eventually about as far below the threshold as it was originally above it.

This last observation, as well as the Threshold Theorem, was discovered by Kermack and McKendrick in 1927. A more precise version of this threshold result is possible by an

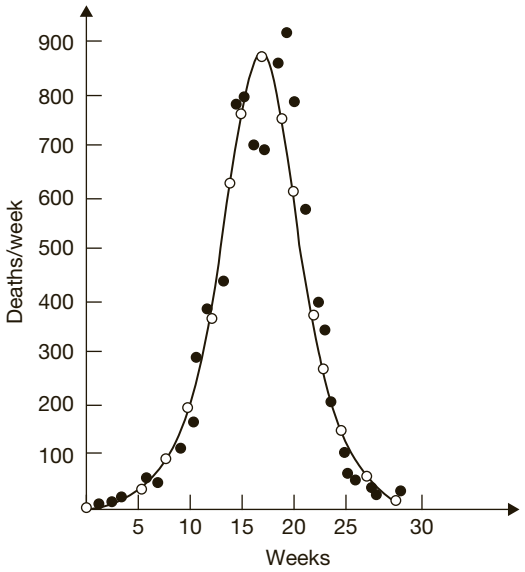


FIGURE 14.8 Kermack and McKendrick's comparison of the predicted curve of dR/dt as a function of t and data on the number of deaths from plague (solid dots) in Bombay over the period of December 17, 1905, to July 21, 1906. The calculated curve conforms roughly to observed figures.

exact solution of the differential equation (41), but the epidemiological implications are quite similar. In addition to the results presented here, Kermack and McKendrick compared their predicted results to an actual epidemic, an outbreak of plague in Bombay in 1905–06. This comparison is shown in Fig. 14.8. The vertical axis represents the number of deaths per week and the horizontal units of measure is time in weeks. Since almost all cases terminated fatally, the vertical component is approximately dR/dr.

E. A Discrete Version of the More General Epidemic Model

We discuss very briefly in this section two discrete analogues of the general epidemic model. We encourage you to explore these discrete models in more detail.

First, the most straightforward translation of the Kermack-McKendrick model to a discrete version is the dynamical system

$$S_{k+1} - S_k = -\beta S_k I_k \tag{45.1}$$

$$I_{k+1} - I_k = \beta S_k I_k - rI_k \tag{45.2}$$

$$R_{k+1} - R_k = rI_k \tag{45.3}$$

where β and r are suitably chosen proportionality constants. With these difference equations, it is possible that the number of infectives could exceed the total population N and the number of susceptibles could become negative. To avoid such complications, we modify the first two equations so they have the form

$$S_{k+1} = \text{maximum}\,(0, S_k + \beta S_k I_k) \tag{46}$$

$$I_{k+1} = I_k - rI_k + \text{minimum}\,(S_k, \beta S_k I_k) \text{ with } 0 < r < 1 \tag{47}$$

Frank De Hoog, Joseph Gani and David Gates [1979] study this model in detail and derive an analogue of the Kermack-McKendrick Theorem.

Second, to study influenza epidemics in England and Wales, Clive Spicer [1979] developed a discrete model using a variable Y_k, the number of newly infective individuals at period k, and also considering p_j, the proportion of the new infectives on any given day k who remained in the population j days later. A Russian scientist O. V. Baroyan had estimated empirically the values of p_j obtaining the following values:

J	0	1	2	3	4	5	6
p_j	1.0	0.9	0.55	0.3	0.15	0.05	0

From this table, we see, for example, that two days after becoming infective, only 55% of these individuals are still infective. No person remains infective for more than five days.

The epidemic begins with S_0 susceptibles and Y_0 new infectives at day 0. The equations for the discrete model are

$$S_{k+1} - S_k = -Y_{k+1}$$

$$Y_{k+1} = \beta S_k$$

$$I_k = \sum_{j=0}^{k} p_j Y_{k-j} = p_0 Y_k + p_1 Y_{k-1} + p_2 Y_{k-2} + \cdots + p_k Y_0$$

$$R_{k+1} = I_k - (I_{k+1} - Y_{k+1})$$

Table 14.3 shows the relations for the first several days of an epidemic between the number of susceptibles, new cases, and total number of infectives.

In Fig. 14.9, we show the number of infectives for the first 30 days of an epidemic that begins with 50 infectives in a population of 1,000.

Table 14.3 The Progress of an Influenza Epidemic

k	Total Number of Infectives I_k	Susceptibles S_k	New Cases Y_k
0	$I_0 = p_0 Y_0$	S_0	Y_0
1	$I_1 = p_0 Y_1 + p_1 Y_0$	$S_1 = S_0 - Y_1$	$Y_1 = \beta S_0 Y_0$
2	$I_2 = p_0 Y_2 + p_1 Y_1 + p_2 Y_0$	$S_2 = S_1 - Y_2$	$Y_2 = \beta S_1 Y_1$
3	$I_3 = p_0 Y_3 + p_1 Y_2 + p_2 Y_1 + p_3 Y_0$	$S_3 = S_2 - Y_3$	$Y_3 = \beta S_2 Y_2$
.	.	.	.
.	.	.	.
.	.	.	.
k	$I_k = \sum_{j=0}^{k} p_j Y_{k-j} = p_0 Y_k + p_1 Y_{k-1} + p_2 Y_{k-2} + \ldots + p_k Y_0$	$S_k = S_{k-1} - Y_k$	$Y_k = \beta S_{k-1} Y_{k-1}$

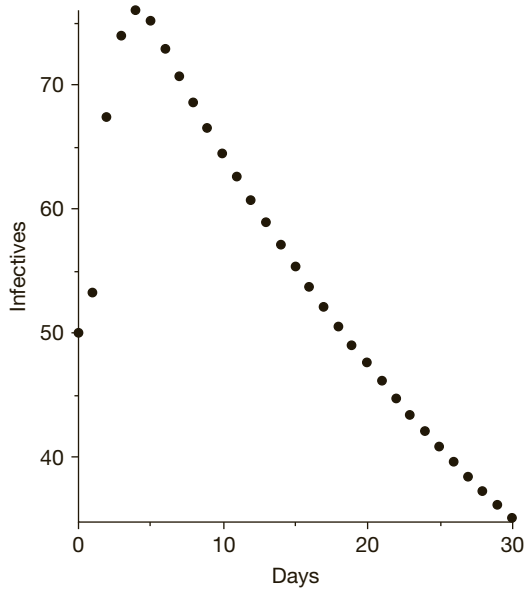


FIGURE 14.9 An instance of the Spicer model, with $S_0 = 950$, $I_0 = 50$, and $\beta = .03$ showing the number of infectives for the first 30 days of the epidemic.

F. Rumors

Variations of mathematical models proven effective in the study of traditional epidemics also aid our understanding the transmission and spread of other phenomena that resemble, at least in some respects, infections.

We'll look at three variants of the epidemic model that have been developed to gain insight into the spread of rumors, the persistence of urban myths, and the problem of problem drinking. Each makes use of a central assumption of the epidemic models.

The inclusion of the term SI in the differential equations for dS/dt and dI/dt is often justified by an appeal to the **Law of Mass Action**. According to this law, when one population group's size is affected by interaction with another population group, a first attempt to measure effectively the intensity of that interaction can usually be modeled by the frequency of contacts between the groups. That frequency is jointly proportional to the two populations: the product of the sizes of the two disjoint groups gives the number of distinct possible meetings. Not only did we see the Law of Mass Action employed in the epidemic models we have studied so far, but we also saw the central role in played in the Predator-Prey and Competitive Hunter models of Chapter 4.

The Law of Mass Action also finds its way into our models of the rumors, urban legends, and problem drinking.

G. Rumors

A rumor is an unverified account or explanation of events circulating from person to person and pertaining to an object, event, or issue of public concern. Rumors may be true or false, but they generally spread rapidly by word of mouth. One person tells a second person, who in turn passes on the content to a third, and the chain continues . . .

We consider a community that we partition into three distinct groups. The first are those who have not yet heard the rumor; they are some times called the "ignorants"; we'll use x (for "unawares" or "unknowers" since x is often used as a symbol for an unknown) to denote how many there are. The second group are those who are actively spreading the rumors, the gossips or "spreaders" of the juicy story. We'll use variable y (for "yenta," the Yiddish word for gossiper). The final group are the people who know the rumor, but have stopped spreading it. These individuals are often described as "stiflers" or "squelchers." We'll denote the size of this group by z (for "zappers").

Now the variables x, y, and z change over time t; we want to model their dynamic behavior. We assume that a rumor is propagated through the population by contact between unawares and yentas, following the Law of Mass Action. More specifically, we'll assume that whenever a yenta meets another person, the gossiper attempts to "infect the mind" of the other by relating the rumor. The other person might be an unaware, a yenta or a stifler. In the first case, the unaware is transformed into a spreader. In the other cases, one or both of the people learns that the rumor is known and so decides to stop telling it.

The number of distinct possible meetings between unawares and yentas is xy and between yentas and zappers is yz. The number of distinct possible meetings of a pair of yentas is $\frac{y(y-1)}{2}$. When an unaware meets a yenta, the outcome is one fewer unaware and one more yenta. When a yenta encounters a zapper, the result is one fewer yenta and one more zapper, but when two yentas stop to talk, the conversation ends with two fewer yentas and two more zappers.

For simplicity, we will assume the size of the population is fixed and that initially there are N people who do not know the rumor and there is one person spreading it. The total population has size $N+1$.

Our model is a linked system of differential equations whose two principal relationships are mathematical representations of our verbal assumptions:

$$\begin{aligned} dx/dt &= (-1)xy \\ dy/dt &= (+1)xy + (-2)\frac{y(y-1)}{2} + (-1)yz = y(x-y+1-z). \end{aligned} \tag{48}$$

with initial conditions $x(0) = N$, $y(0) = 1$

Since $x + y + z = N + 1$ for all time t, we have $x - y + 1 - z = 2x - N$. Thus, we can write our pair of differential equations as

$$dx/dt = -xy \tag{49.1}$$

$$dy/dt = y(2x - N) \tag{49.2}$$

From this pair of equations, we obtain

$$\frac{dy}{dx} = \frac{\frac{dy}{dt}}{\frac{dx}{dt}} = \frac{y(2x-N)}{-xy} = -2 + \frac{N}{x} \tag{50}$$

The solution to the differential equation

$$\frac{dy}{dx} = -2 + \frac{N}{x} \tag{51}$$

with initial condition $y = 1$ when $x = N$ is

$$2x(t) + y(t) + N\ln\frac{N}{x(t)} = 2N + 1 \text{ for all } t > 0 \tag{52}$$

Let U_N denote the long-term proportion of the population that never hears the rumor—that is, $U_N = lim_{t\to\infty}\frac{x(t)}{N}$. Then Eq. (52) and the fact that $lim_{t\to\infty} y(t) = 0$ gives us

$$2U_N - \ln U_N = 2 + \frac{1}{N} \tag{53}$$

As the overall population $N + 1$ increases, $lim_{N\to\infty} U_N = U$ satisfies

$$2U - \ln U = 2 \text{ or } 2(1 - U) + \ln U = 0 \tag{54}$$

You can think of the value of U as the long-term proportion of a very large population who never hear the rumor. There are two solutions of this equation: $U = 1$ and $U \sim .2032$. With the given initial condition, x is always decreasing, so we cannot have $U = 1$. Thus, under the assumptions of our model, if the population is large, then about 20% of the people will never learn the rumor.

Note that if we write Eq. (51) as $\frac{dy}{dx} = 2\left(-1 + \frac{N/2}{x}\right)$, then it has essentially the same structure as Eq. (30) of the more general deterministic epidemic model $\frac{dI}{dS} = \left(-1 + \frac{\rho}{S}\right)$ with a relative removal rate of $N/2$. Thus, we can use the techniques we applied to Eq. (30) to gain further insight into the rumor model.

H. Persistence of Urban Legends

You've probably heard at least one of these stories:

- "Giant albino alligators live deep in the sewers under New York city; they're descendants of small, pet gators city residents picked up while touristing in Florida and then flushed down their toilets when the animals became too difficult to handle in a Manhattan apartment."
- "Soft bubble gum has a secret ingredient that keeps it chewy: spider eggs."
- "Cats can suck the air out of babies. A registered nurse says that cats get in the bed with babies and lick their mouths and then suck the air out of them and the babies die."
- "A man was on a business trip alone, and went out to a bar one night to have a cocktail. He woke up the next morning in an unfamiliar hotel room with severe pain in his lower back. He was taken to the emergency room, where doctors determined that, unknown to him, he had undergone major surgery the night before. One of his kidneys had been removed, cleanly and professionally. He was the victim of a crime ring that drugs out-of-town visitors, surgically removes organs from their bodies, and sells the organs on the black market."

These stories are all examples of "Urban Legends" or "Urban Myths": a special class of rumors that are false, but persistent short tales spread classically by word of mouth but now more frequently by email. Sociologist Andrew Noymer noted that persistence is "what sets urban legends apart from rumors more generally, which may disappear almost as soon as they arise." Noymer [2001] developed several mathematical approaches based on epidemic models to examine how urban legends may become entrenched and continue in circulation for long periods, even though there may be skeptics who actively try to convince others that the legends are untrue. We will examine here one variation of Noymer's models. Our model is conceptually the same as his, but Noymer casts it in terms of partial differential equations; we will use a system of discrete difference equations with each step corresponding to a new week.

Noymer notes a strong analogy between epidemic models and rumor diffusion models. Taking measles as a representative infectious disease, Noymer points out that

> *Measles is highly contagious, and is spread by infected-to-susceptible contact. . . . Rumors are also highly contagious: what differentiates rumors from other pieces of information is that the possessor of a rumor has an irresistible urge to tell others . . .*
>
> *"Belief in a rumor and desire to spread the rumor are here taken to be identical, though in practice belief may persist even after the burning desire to spread a new rumor wanes. The contact spread of pathogens and the contact spread of rumors is analogous . . .*
>
> *"In two respects, the measles-rumors analogy breaks down. Measles has a latent period which is unlike most rumors; with rumors there is no distinction between infection and contagiousness. Measles involves recovery (or death) within a few weeks of initial infection, whereas some rumors may be believed for years. These differences are easy to deal with from the modeling perspective."*

In building his models, Noymer begins by adding a fourth state to the classic three-state SIR Kermack-McKendrick model. The new state is made up of those who do not understand the urban legend. These are very young babies and children. They have immunity from transmission of the rumor. Borrowing from the medical literature that often designates this group as those "protected by maternal antibodies," Noymer denotes this population of toddlers as M.

Fig. 14.10 shows the flow of individuals through the states of the model. Movement through the states depends not only on time but also the age of individuals; hence, Noymer's models are described as "age-structured" ones.

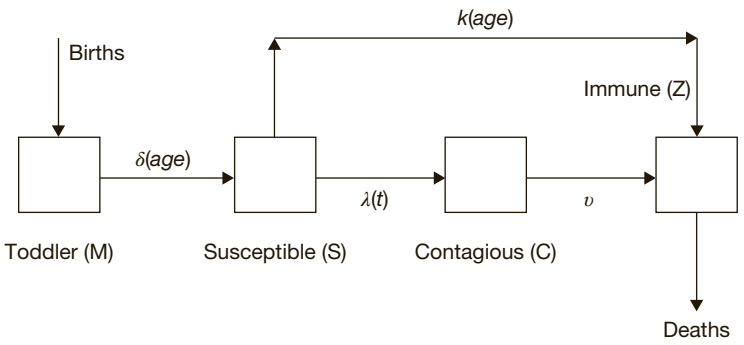


FIGURE 14.10 Flow diagram of Noymer model.

Toddlers may move from the immune state into the susceptible state, but the rate of movement depends on the age of the toddlers. Noymer posits that children below the age of three years (156 weeks) are too young to comprehend the urban legend. Thus, we can model the change in population of toddlers as

$$\begin{aligned} &M_{k+1} - M_k = \text{Births} - \delta(age)M_k \\ &\text{where} \\ &\delta(age) = \begin{cases} 0 & \text{if } age < 156 \\ .0064 & \text{if } age \geq 156 \end{cases} \end{aligned} \tag{55}$$

Note that $.0064 = 1/156$.

Once in the Susceptible (S) state, a certain fraction may become contagious by the usual Law of Mass Action factor, but another fraction may become so skeptical that they will not believe a rumor. Noymer assumes that between the ages of 3 and 6, there is rapid recruitment into the contagious state, but after 6 some children are "savvy enough" and "will not believe everything they are told."

The change in the susceptible population from one week to the next is then given by

$$\begin{aligned} &S_{k+1} - S_k = \delta(age)M_k - k(age)S_k - \frac{\lambda}{N}S_k C_k \\ &\text{where} \\ &k(age) = \begin{cases} 0 & \text{if } age < 312 \\ .0014 & \text{if } age \geq 312 \end{cases} \end{aligned} \tag{56}$$

and λ is a per-capita transmission rate.

Finally, we examine the change in population of those spreading the urban legend and those who have recovered from the belief that it might be true. We assume that there is some constant recovery rate υ per person so that

$$C_{k+1} - C_k = \frac{\lambda}{N}S_k C_k - \upsilon C_k \tag{57}$$

and

$$Z_{k+1} - Z_k = k(age)S_k + \upsilon C_k - Deaths \tag{58}$$

Noymer assumes a population of fixed size N where the number of births per week equals the number of deaths. He also assumes for simplicity that the population is in equilibrium as far as age structure is concerned.

Fig. 14.11 and Fig 14.12 show the results of a typical run of the model. Here the parameter values are $\upsilon = 0.04/\text{week}$ and $\lambda = .20208/\text{week}$. The population size was fixed at 100,000 with 100 births and 100 deaths per week. We began with only three contagious individuals and 1,000 susceptibles.

Note that there is a major growth in the contagious population when our large initial group of toddlers turns 3 years old. Although most of them have become skeptics by their late teens, enough remain believers in the legend that they can infect the next generations of

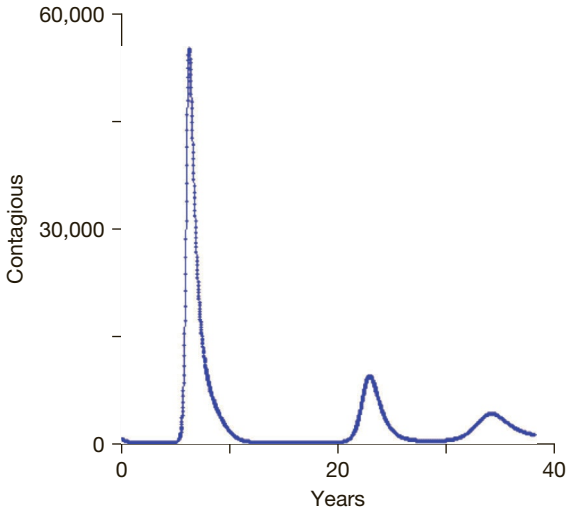


FIGURE 14.11 Persistence of an urban legend for Noymer's model.

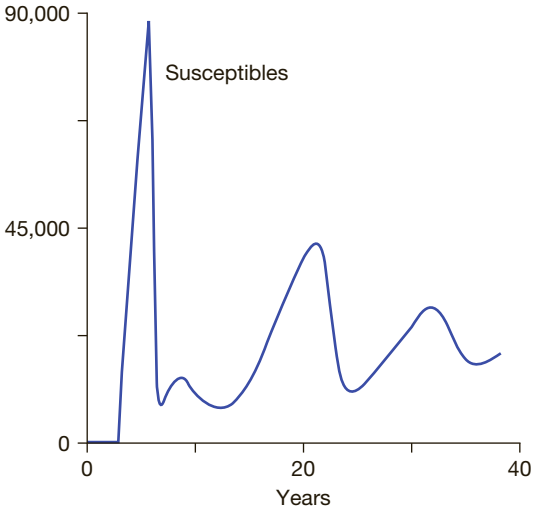


FIGURE 14.12 Fluctuations in the susceptible population over time in Noymer's model.

toddlers when they become susceptibles. There is a second and then a third spurt, so we see recurrent waves of activity as the urban legend is passed on and persists over a long period of time.

One weakness of this model is the assumption of a constant rate v of "recovery." A constant rate makes more sense for a disease such as measles than for belief in an urban legend. "After all," Noymer argues, "if someone believes a rumor in the first place, why should she spontaneously stop believing the rumor?"

For our second model, we incorporate Noymer's suggestion: "Suppose instead that the rumor is believed indefinitely until it is challenged through contact with skeptics." With this assumption, conversion from contagious to skeptic is proportional to the number of

interactions between the these two groups. Once again, we'll use the Law of Mass Action to replace Eq. (57) with

$$C_{k+1} - C_k = \frac{\lambda}{N} S_k C_k - \frac{\gamma}{N} C_k Z_k \tag{59}$$

For ease in testing different assumptions about the relative success of convincing a susceptible to believe and convincing a believer to turn skeptical, we rewrite this last equation as

$$C_{k+1} - C_k = \frac{\lambda}{N} S_k C_k - \frac{q\lambda}{N} C_k Z_k \tag{60}$$

The second model incorporates Noymer's idea that "skeptics transmit their immunity to the contagious in the same fashion that the contagious transmit the rumor to the susceptible."

To summarize, our second model is the system of difference equations

$$M_{k+1} - M_k = births - \delta(age) M_k \tag{61}$$

$$S_{k+1} - S_k = \delta(age) M_k - \kappa(age) S_k - \frac{\lambda}{N} S_k C_k \tag{62}$$

$$C_{k+1} - C_k = \frac{\lambda}{N} S_k C_k - \frac{q\lambda}{N} C_k Z_k \tag{63}$$

$$Z_{k+1} - Z_k = \kappa(age) S_k + \frac{q\lambda}{N} C_k Z_k. - deaths \tag{64}$$

Noymer investigates two qualitatively different cases for the second model. In the first, those believing the rumor are fairly ready to change their mind $(q = 0.3)$ and in the second, "those believing the rumor are loath to be skeptical" $(q = 0.01)$.

Fig. 14.13 shows graphs of the proportion of the population that remains infected under these two choices for q. In the case $q = 0.01$, there is an initial rapid increase to over 30% of the population, then a gradual decline to a level remaining constant at about 5%. When $q = 0.3$ and rumor believers are more ready to change their belief, we see

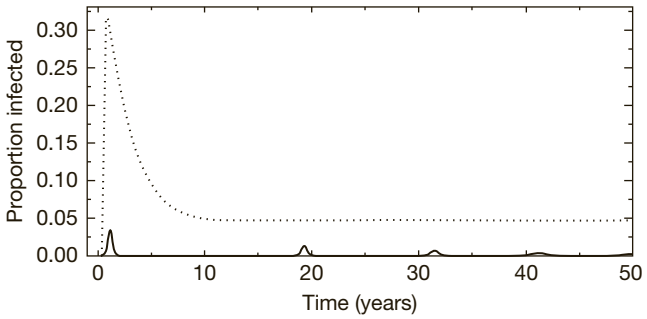


FIGURE 14.13 Noymer's skeptic model showing proportion of the population that is "infected" over a long time period. The dotted curve corresponds to $q = 0.01$ and the solid curve is the case $q = 0.3$.

that not as many people believe the rumor at any one time, but there are short-term bumps in the numbers at 10−20-year gaps.

I. Problem Drinking

As we have seen, epidemiological models can be used to study the dynamics of the transmission of infectious diseases, short-lived rumors, and urban legends. Modelers have also used variations on these approaches to examine social and behavioral processes such as violence, eating disorders, and drug addictions.

Fabio Sánchez and his colleagues [2007] employed an epidemic-type model to examine the dynamics of drinking alcoholic beverages. As they observe,

> *There are clearly differences in the generation of addictive behaviors and the transmission of infectious diseases. However, the fact remains that the acquisition of both can be modeled . . . as the likely result of contacts between individuals in given environments. For example, the development of alcohol use among young people and the influence of 'supportive environments' on the development and maintenance of heavy drinking, alcohol abuse, dependence and problems among adults, are predicated upon the combined effects of social influence and access to alcohol. Thus, additional understanding of the dynamics of drinking behaviors may result from the use of a perspective that models drinking as the result of contacts of susceptibles with individuals in distinct drinking states."*

Sánchez et al. model problem drinking as an acquired state that is the result of frequent or intense interaction between individuals in three drinking states. They partition a fixed-sized population of size N into three groups: the occasional and moderate drinkers who function in the role of susceptibles, the problem drinkers or instigators who promote a culture of drinking, and the temporarily recovered. The variables S, I, and R in this model refer to the percentage of individuals in each of these groups. Thus, $S + I + R = 1$. We are assuming that the time scale of interest is short enough that the overall population size does not change significantly. New community members join the population as moderate drinkers and mix at random with the remainder of the population.

Fig. 14.14 shows a flow model for the dynamics of this model. New susceptibles join the environment at a constant rate. Some of these leave the environment, and others may be transformed into problem drinkers. Some of the problem drinkers also leave the environment while others may become temporarily recovered. Some of the temporarily recovered also leave the scene, but some can relapse into problem drinkers.

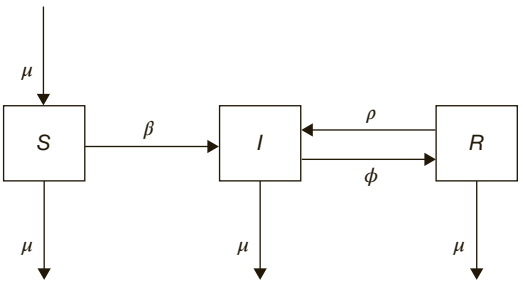


FIGURE 14.14 Flow diagram of the Sánchez model for the contagious drinking problem.

The differential equations for our model are

$$\frac{dS}{dt} = \mu - \beta SI - \mu S \tag{65.1}$$

$$\frac{dI}{dt} = \beta SI + \rho RI - (\mu + \phi)I \tag{65.2}$$

$$\frac{dR}{dt} = \phi I - \rho RI - \mu R \tag{65.3}$$

$$S + I + R = 1 \tag{65.4}$$

Two features distinguish the Sánchez model from the SIR model we examined earlier in this chapter. The major difference is that movement into the R group need not be permanent; the instigators may well corrupt a temporarily recovered person and cause him to relapse into a problem drinker. The other difference is the assumption that individuals may leave the system from any state.

Our analysis of this model begins by considering the situation in which there is no significant drinking issue—that is, problem drinkers are so rare that their numbers are insignificant, and no recovery treatments are available, so the R state is not present.

Epidemiologists use the term *Basic Reproductive Number* for the expected (average) number of new infectious cases in a completely susceptible population produced directly by a single case during its entire period of infectiousness. When the Basic Reproductive Number is 1, each case only reproduces itself so the number of cases stays steady, neither growing nor shrinking. A higher Basic Reproductive Number represents a more transmissible disease, one that can more broadly spread in the population.

Let $\mathfrak{R}_0 = \frac{\beta}{\mu}$ the product of the transmission rate β and the average time, $\frac{1}{\mu}$, an individual remains a problem drinker. Then $\mathfrak{R}_0$ is the number of secondary cases generated by a typical problem drinker in the no significant drinking environment. It is the Basic Reproductive Number in this environment.

In the absence of the R state, the equation for the rate of change of I in our model becomes

$$\frac{dI}{dt} = \beta SI - \mu I = I(\beta S - \mu) \tag{66}$$

If $\mathfrak{R}_0 = \frac{\beta}{\mu} < 1$, then $\beta < \mu$, and since S is the proportion of susceptibles in the population, we have $0 < S < 1$, so that $\beta S < \beta < \mu$, and hence $\beta S - \mu < 0$, so I will decline to 0 as $t \to \infty$. On the other hand, if $\mathfrak{R}_0 > 1$, then even the introduction of a single problem drinker will cause I to approach some positive value I^* as time advances.

In the model *with recovery*, a problem drinker remains one for an average time of $\frac{1}{\mu + \varphi}$ but still has the transmission rate β so the basic reproductive number with recovery, $\mathfrak{R}_\phi$ is $\frac{\beta}{\mu + \varphi}$. The epidemic will not necessarily die out if $\mathfrak{R}_\phi < 1$ but if $\mathfrak{R}_\phi > 1$, there will be a long-term persistent population of problem drinkers.

We introduce one more threshold here, $\mathfrak{R}_\rho = \frac{\rho}{\mu+\varphi}$, which measures the average number of temporarily recovered individuals a single problem drinker causes to relapse. The number $\mathfrak{R}_\rho$ may be called the "Basic Reproductive Number with Treatment."

We continue our analysis, as we did for the Richardson arms race model and the interactive species models, by examining equilibrium points, the points at which all the time derivatives are simultaneously zero. One such point is $(S, I, R) = (1.0, 0.0, 0.0)$. We want to determine when there are critical points (S^*, I^*, R^*) that have all components positive. Such a critical point would be a solution where problem drinking may become established.

We'll begin with the condition that $dS/dt = 0$:

$$\mu - \beta SI - \mu S = 0$$

which we may rewrite as

$$S = \frac{\mu}{\beta I + \mu}.$$

If we examine the condition $dI/dt = 0$, then we can obtain another relation between S and I at equilibrium:

$$\beta SI + \rho RI - (\mu + \phi)I = 0$$

or

$$\beta S + \rho R - (\mu + \phi)I = 0$$

but at a positive equilibrium point, $I > 0$, so

$$\beta S + \rho R - (\mu + \phi) = 0$$

but $S + I + R$ is always equal to 1, so

$$\beta S + \rho(1 - S - I) - (\mu + \phi) = 0.$$

If we solve this equation for S in terms of I, we find

$$S = \frac{\mu + \varphi - \rho + \rho I}{\beta - \rho}.$$

Thus, at equilibrium, we would have

$$\frac{\mu}{\beta I + \mu} = S = \frac{\mu + \varphi - \rho + \rho I}{\beta - \rho}$$

so

$$\mu(\beta - \rho) = (\mu + \varphi - \rho + \rho I)(\beta I + \mu).$$

Multiplying out and collecting terms results in

$$(\rho\beta)I^2 - [\rho(\beta - \mu) - \beta(\mu + \varphi)]I + (\mu^2 + \mu\varphi - \mu\beta) = 0$$

which we may rewrite, after some algebraic manipulation, as

$$I^2 - \left[1 - \frac{\mu}{\beta} - \frac{\mu + \varphi}{\rho}\right]I + \frac{\mu}{\beta}\left[\frac{\mu + \varphi}{\rho} - \frac{\beta}{\rho}\right] = 0.$$

Using our symbols for the basic reproductive numbers, we may write this quadratic equation as

$$I^2 - [1 - 1/\mathfrak{R}_0 - 1/\mathfrak{R}_\rho]I + 1/\mathfrak{R}_0[1/\mathfrak{R}_\rho - \beta/\rho] = 0.$$

Now this quadratic equation has the form

$$I^2 - BI + C = 0$$

where $B = 1 - 1/\mathfrak{R}_0 - 1/\mathfrak{R}_\rho$ and $C = (1/\mathfrak{R}_0)[1/\mathfrak{R}_\rho - \beta/\mu]$.

The quadratic equation has two distinct roots between 0 and 1 whenever $0 < B < 1$, $C > 0$, and $B^2 - 4C > 0$. [See Exercise 46.]

Let's see when these inequalities hold:

$C > 0$: C is the product of two factors. The first one, $1/\mathfrak{R}_0$, is positive, so C is positive exactly when $\left[1/\mathfrak{R}_\rho - \beta/\rho\right] = \left[\frac{\mu+\varphi}{\rho} - \frac{\beta}{\rho}\right] = \left[\frac{\mu+\varphi-\beta}{\rho}\right] > 0$, which occurs only when $\mu + \varphi > \beta$—that is, $\frac{\beta}{\mu+\varphi} < 1$. But $\frac{\beta}{\mu+\varphi} = \mathfrak{R}_\phi$. Thus, $C > 0$ whenever $\mathfrak{R}_\phi < 1$.

$B^2 - 4C > 0$: If the discriminant of the quadratic equation is positive, then we must have $\mathfrak{R}_\rho > 1$ and $0 < \mathfrak{R}_c < \mathfrak{R}_\phi < 1$ where

$$\mathfrak{R}_c = \frac{\rho}{\beta}\left[\frac{1}{1 + \frac{1}{\mathfrak{R}_0}} - 2\sqrt{\frac{1}{\mathfrak{R}_0} - \frac{\mu}{\rho}}\right]$$

Sánchez et al. investigated four separate "thresholds" affecting the qualitative behavior of the model. Table 14.4 describes these thresholds.

Table 14.4 Description of Threshold Conditions

Thresholds	Description
$\mathfrak{R}_0$	Number of secondary cases generated by a "typical" problem drinker in a nondrinking population
$\mathfrak{R}_\phi$	Basic reproductive number with recovery
$\mathfrak{R}_\rho$	Number of secondary cases generated by a "typical" problem drinker in a population of temporarily recovered individuals
$\mathfrak{R}_c$	Critical value to where drinking communities can be under control

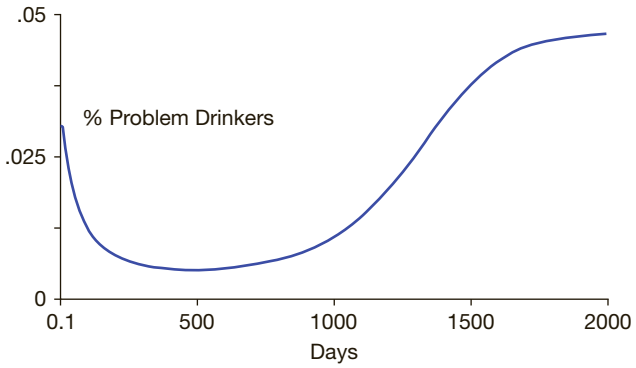


FIGURE 14.15 A sufficiently high initial population of problem drinkers can, after an initial decline, ultimately climb to a stable level larger than at the beginning of the epidemic.

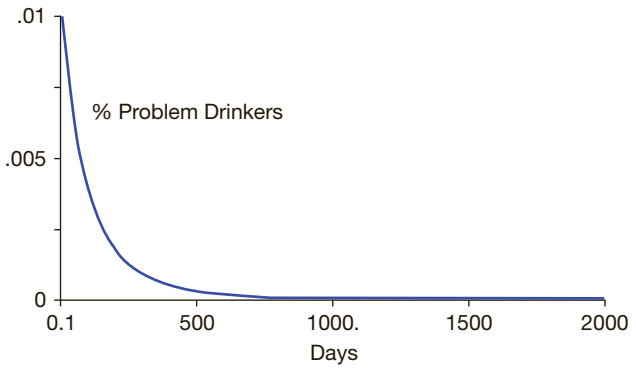


FIGURE 14.16 The problem drinker population can be reduced to 0 if it is not too high initially.

Figs. 14.15 and 14.16 display two different scenarios for solutions of the system of differential equations presented in Eqs. (65.1)−(65.4) showing the evolution of the proportion of the population that are serious drinkers over an extended time period. The parameter values in both cases are $\mu = .0000548$, $\beta = 0.19$, $\phi = 0.2$, and $\rho = 0.21$. Fig. 14.14 shows the situation in which the initial conditions are $s_o = 0.97$, $d_o = .03$, and $r_o = 0$. Fig. 14.15 is the case in which $s_o = .99$, $d_o = 0.01$, and $r_0 = 0$. In both cases there is an initial drop in the problem drinkers. In the second case, in which there is smaller initial proportion of problem drinkers, their population dies out over time. If their initial population is slightly higher (3% vs. 1%), an initial descent ultimately reverses and the long-term proportion tends to a level higher than at the start.

In a subsequent paper, Ariel Cintróán-Arias, Fabio Sánchez, and others review this model and present more sophisticated probabilistic models using stochastic processes and Markov Chains; see Cintróán-Arias [2009] for details.

J. The Mickens Model: Square Root Dynamics

In formulating a mathematical model of the spread of a communicable disease, we must consider the rate at which uninfected persons acquire the illness. There must be some form of contact between a susceptible and an infective that leads to the transmission of

the disease. Not every contact results in passing the disease, but the more contacts there are, the more likely it is that a previously healthy person will contract the illness. The total number of contacts between susceptibles and infectives will be an important quantity. The quantity of contacts should depend on the population S of susceptibles and the population I of infectives. If both S and I are large, the number of contacts will be large. If S and I are small, there will be few contacts and fewer opportunities for the passage of the disease.

The classic SIR model (often called the Kermack-McKendrick model) we studied earlier in this chapter incorporates the Law of Mass Action to measure the number of contacts. That principle states that the number of contacts is jointly proportional to the two populations—that is, there is a positive constant $\beta < 1$ such that the rate of transmission is βSI. But there are other quantities involving S and I that grow as both S and I increase. One such measure replaces the product of the populations with product of their square roots—that is, the rate of transmission is $\beta\sqrt{S}\sqrt{I}$.

In this section, we will investigate a model, due to Ronald Mickens, that uses this idea. The Mickens [2012] model is the system of differential equations

$$\frac{dS}{dt} = -\beta\sqrt{S}\sqrt{I}, \quad \beta > 0 \tag{75.1}$$

$$\frac{dI}{dt} = \beta\sqrt{S}\sqrt{I} - r\sqrt{I}, \quad r > 0 \tag{75.2}$$

$$\frac{dR}{dt} = r\sqrt{I} \tag{75.3}$$

with initial conditions

$$S(0) = S_0, \quad I(0) = I_0 > 0, \quad R(0) = R_0 = 0$$

and the relation

$$S(t) + I(t) + R(t) = N \text{ for all } t \geq 0.$$

We shall see that the Mickens model exhibits much of the same qualitative behavior as the classic SIR model. It satisfies a Threshold Theorem, for example. It also makes a more realistic prediction about the long-term behavior of the infective population in the case in which are no susceptibles (see Exercise 50). Most important, perhaps, it is possible to find explicit solutions for I, S, and R in terms of elementary functions of t, a feature that the classic model lacked.

Since the susceptible and infective populations can't be negative, the first observation we make is that Eq. (75.1) implies that S is always decreasing and Eq. (75.3) implies that R is always increasing. From Eq. (75.2), we see that the dI/dt is positive only for $S > \frac{r^2}{\beta^2}$ so that the infective population is on the increase whenever S exceeds $\frac{r^2}{\beta^2}$ and is decreasing as a function of time when the number of susceptibles drops below $\frac{r^2}{\beta^2}$.

Let's examine first the relationship between S and I. From Eqs. (75.1) and (75.2), we have

$$\frac{dI}{dS} = \frac{\beta\sqrt{S}\sqrt{I} - r\sqrt{I}}{-\beta\sqrt{S}\sqrt{I}} = -\frac{\beta}{\beta}\frac{\sqrt{I}}{\sqrt{I}}\frac{\sqrt{S} - \frac{r}{\beta}}{\sqrt{S}} = \frac{\frac{r}{\beta} - \sqrt{S}}{\sqrt{S}} = \frac{r}{\beta}S^{-\frac{1}{2}} - 1 \tag{76}$$

Separating the variables and integrating yields

$$\int 1dI = \int \left(\frac{r}{\beta}S^{-\frac{1}{2}} - 1\right)dS$$

and hence

$$I = 2\frac{r}{\beta}S^{\frac{1}{2}} - S + C = 2\frac{r}{\beta}\sqrt{S} - S + C \tag{77}$$

for some constant C. Letting the positive constant $r/\beta = k$, we see that the number of infectives I is a function of the number of susceptibles S of the form

$$I = f(S) = 2k\sqrt{S} - S + C \tag{78}$$

The derivatives of this function are

$$f'(S) = \frac{k}{\sqrt{S}} - 1 \text{ and} f''(S) = \frac{-k}{2S^{\frac{3}{2}}} \tag{79}$$

Since the second derivative is negative, the graph of I as a function of S will be concave down with a maximum value where the first derivative is zero. Now $f'(S) = 0$ when $S = k^2 = \frac{r^2}{\beta^2}$. Furthermore, $f'(S) > 0$ when $S < \frac{r^2}{\beta^2}$ and $f'(S) < 0$ when $S > \frac{r^2}{\beta^2}$. In the Mickens model, S is always decreasing. Thus, if S_0 exceeds $\frac{r^2}{\beta^2}$, as S decreases, I will initially increase. When S drops below $\frac{r^2}{\beta^2}$ and keeps decreasing, I will decrease. Thus, we have a *Threshold Theorem* for the Mickens model: There will be an epidemic if and only the initial population of susceptibles exceeds $\frac{r^2}{\beta^2}$.

Fig. 14.17 shows two typical trajectories. The initial conditions $(S^{\#}, I^{\#})$ do not lead to an epidemic while the initial conditions $(S^{\#\#}, I^{\#\#})$ do. The symbols $S_\infty^{\#}$ and $S_\infty^{\#\#}$ denote the remaining susceptible populations at the end of the disease spreading process—that is, $S_\infty = \lim_{t\to\infty} S(t)$.

Returning to our solution, Eq. (77), of the differential equation for dI/dS,

$$I = 2\frac{r}{\beta}\sqrt{S} - S + C$$

we can evaluate the constant C by using the initial numbers of infectives and susceptibles at time $t = 0$:

$$I_0 = 2\frac{r}{\beta}\sqrt{S_0} - S_0 + C \text{ so } C = I_0 + S_0 - 2\frac{r}{\beta}\sqrt{S_0} \tag{80}$$

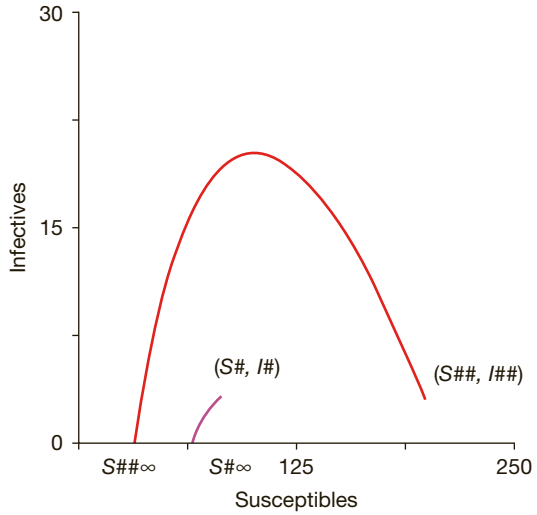


FIGURE 14.17 Trajectories for the Mickens model in the $S-I$ plane. Here $\beta = .03$, $r = .3$ and $I_0 = 11$. The lower curve represents the trajectory if $S_0 = 80$ and the upper curve corresponds to $S_0 = 200$. In the latter case, there is an epidemic as the number of infectives initially increases. In the former case, there is no epidemic; the number of infectives strictly decreases at all times.

and hence

$$I(t) + S(t) - 2\frac{r}{\beta}\sqrt{S(t)} = I_0 + S_0 - 2\frac{r}{\beta}\sqrt{S_0} \tag{81}$$

As we have seen, the parameter $\frac{r^2}{\beta^2}$ plays a key role in analyzing the dynamics of the Mickens model. We will use S* to denote $\frac{r^2}{\beta^2}$. Thus, we can write the first-integral equation as

$$I(t) + S(t) - 2\sqrt{S^* S(t)} = I_0 + S_0 - 2\sqrt{S^* S_0} \tag{82}$$

Our condition for the existence of an epidemic is $S_0 > S^*$—that is, the ratio $R_0 = S_0/S^*$ must exceed 1. The ratio R_0 is called the *basic reproduction number*. Note that the disease will die out if $0 < R_0 < 1$ while an epidemic will occur if $R_0 > 1$. We can interpret R_0 as the number of additional infections induced into a susceptible population by a single infected individual.

Suppose we do have $S_0 > S^*$, so an epidemic does occur. What is the largest number I_{max} of infective people we will ever see? As we noted above, I reaches its maximum value when $S = \frac{r^2}{\beta^2} = S^*$. Substituting into Eq. (82), we have

$$I_{max} + S^* - 2\sqrt{S^* S^*} = I_0 + S_0 - 2\sqrt{S^* S_0}$$

or

$$I_{max} = I_0 + S_0 - 2\sqrt{S^* S_0} + S^* = I_0 + \left(S_0 - \sqrt{S^*}\right)^2 \tag{83}$$

Example 1

For our model with $\beta = .03$, $r = .3$, $S_0 = 225$, and $I_0 = 11$, we have $S^* = 100$ and $I_{max} = 11 + (15-10)^2 = 36$.

When the epidemic has run its course, there will be no more infectives—that is, $I_\infty = \lim_{t\to\infty} I(t) = 0$. We can use this fact and the solution to our differential equation for dI/dS to determine how many susceptibles never succumbed to the disease, the number remaining when the epidemic is over. We seek $S_\infty = \lim_{t\to\infty} S(t)$. We know that $S_\infty < S^* < S_0$. Substituting $I_\infty = 0$ and S_∞ into our solution,

$$I(t) + S(t) - 2\sqrt{S^* S(t)} = I_0 + S_0 - 2\sqrt{S^* S_0}$$

we have

$$0 + S_\infty - 2\sqrt{S^*}\sqrt{S_\infty} = I_0 + S_0 - 2\sqrt{S^* S_0} = C \tag{84}$$

If we let $x = \sqrt{S_\infty}$, then

$$x^2 - 2\sqrt{S^*}x - C = 0 \tag{85}$$

The solution of this quadratic equation is

$$x = \frac{2\sqrt{S^*} \pm \sqrt{4S^* + 4C}}{2} = \sqrt{S^*} \pm \sqrt{S^* + C} \tag{86}$$

and therefore,

$$S_\infty = x^2 = S^* \pm 2\sqrt{S^*}\sqrt{S^* + C} + S^* + C = 2S^* + C \pm 2\sqrt{S^*}\sqrt{S^* + C} \tag{87}$$

Recall that $C = I_0 + S_0 - 2\frac{r}{\beta}\sqrt{S_0} = I_0 + S_0 - 2\sqrt{S^*}\sqrt{S_0}$ so that our formula for S_∞ becomes

$$S_\infty = x^2 = 2S^* + I_0 + S_0 - 2\sqrt{S^*}\sqrt{S_0} \pm 2\sqrt{S^*}\sqrt{S^* + I_0 + S_0 - 2\sqrt{S^*}\sqrt{S_0}} \tag{88}$$

Some algebraic manipulations on this last formula yield a more compact representation as

$$S_\infty = S^*\left[1 - \sqrt{\left(\sqrt{\frac{S_0}{S^*}} - 1\right)^2 + \frac{I_0}{S^*}}\right]^2 \tag{89}$$

where we keep the root of the quadratic equation that gives a value of S_∞ below S^* and S_0. The total number of susceptibles who contracted the disease, S_{ill}, is given by $S_{ill} = S_0 - S_\infty$, and the number of individuals who were ever infected is $I_{total} = I_0 + S_{ill} = I_0 + S_0 - S_\infty$.

Example 2

For our model with $\beta = .03$, $r = .3$, $S_0 = 225$, $I_0 = 11$, and $S^* = 100$, we have $S_\infty = 16$, $S_{ill} = 209$, and $I_{total} = 220$.

One of the nicest features of the Mickens model is that it is possible to obtain exact solutions of the differential equations so that we will have explicit formulas for S, I, and R as functions of t. The technique is to make a change of variables to get an equivalent system of *linear* differential equations.

To begin, let $u = \sqrt{S} = S^{\frac{1}{2}}$ and $v = \sqrt{I} = I^{\frac{1}{2}}$. Then

$$\frac{du}{dt} = \frac{1}{2} S^{\frac{-1}{2}} \frac{dS}{dt} = \frac{1}{2} \frac{1}{\sqrt{S}} (-\beta \sqrt{S} \sqrt{I}) = -\frac{\beta}{2} \sqrt{I} = -\frac{\beta}{2} v = -bv \tag{90}$$

where $b = \frac{\beta}{2}$.

A similar computation shows

$$\frac{dv}{dt} = bu - s \tag{91}$$

where $s = r/2$.

Thus, our new system is the pair of linear differential equations

$$u'(t) = \frac{du}{dt} = -bv \qquad v'(t) = \frac{dv}{dt} = bu - s \tag{92}$$

From this linked pair of equations, we can obtain separate second-order linear differential equations for u and v:

$$u'' = (u')' = (-bv)' = -bv' = -b(bu - s) = -b^2 u + bs$$

so

$$u'' + b^2 u = bs \tag{93}$$

A similar calculation shows that $v'' = -b^2 v$, so that $v'' + b^2 v = 0$. Thus, we have two second-order linear equations

$$u'' + b^2 u = bs \qquad v'' + b^2 v = 0 \tag{94}$$

The initial conditions for these equations are easily computed as

$$\begin{aligned} u(0) = u_0 = \sqrt{S_0} \quad & u'(0) = -b\sqrt{I_0} \\ v(0) = v_0 = \sqrt{I_0} \quad & v'(0) = b\sqrt{S_0} - s \end{aligned}$$

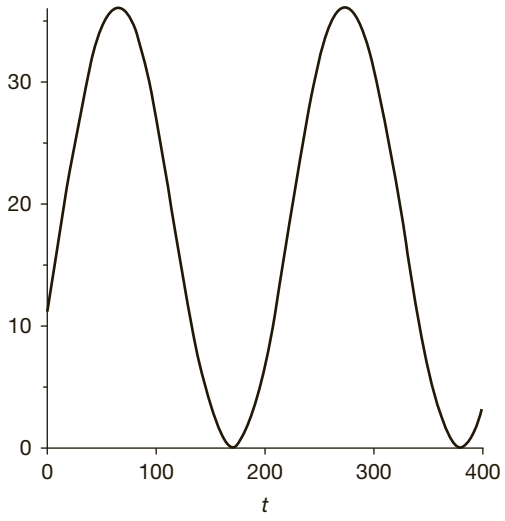


FIGURE 14.18 Oscillations in the graph of $v^2(t)$ that can't occur in a real epidemic.

We can solve each equation independently of the other. Standard techniques (see Appendix V) yield the solutions

$$u(t) = \left[\sqrt{S_0} - \frac{s}{b}\right]\cos(bt) - \sqrt{I_o}\sin(bt) + \frac{s}{b} \tag{95}$$

$$v(t) = \sqrt{I_0}\cos(bt) + \left(\sqrt{S_0} - \frac{s}{b}\right)\sin(bt) \tag{96}$$

Recalling that that $v = \sqrt{I}$ and $u = \sqrt{S}$, it is tempting to set $I = v^2$ and $S = u^2$, but this would not be quite correct. Fig. 14.18 shows the graph of a typical v^2. Note that it has oscillations that a true function for the infective population would not have. We know that if $S_0 < S^*$, then $I(t)$ monotonically decreases to zero, and when $S_0 > S^*$, the function $I(t)$ initially increases to a positive maximum before it declines to zero. When the infective population reaches 0, it remains there forever more. There is a time t_c at which $I(t_c) = 0$ with $I(t) = 0$ for all $t \geq t_c$.

This problem is easily addressed. Note first that the function, which is identically 0, is also a solution to the differential equation $\frac{dI}{dt} = \beta\sqrt{S}\sqrt{I} - r\sqrt{I}$. If we construct a piecewise function for $I(t)$ that is equal to $v(t)^2$ up to $t = t_c$ and equal 0 thereafter, then $I(t)$ will be solution to our differential equation. Similarly, we can let $S(t)$ be equal to $u(t)^2$ up to $t = t_c$ and equal to S_∞ after.

All we have left to do is compute the value t_c where $v(\mathrm{t}_c) = 0$. Now $v(\mathrm{t}_c) = 0$ when $\left(\sqrt{S_0} - \frac{s}{b}\right)\sin(bt_c) = -\sqrt{I_0}\cos(bt_c)$ so that $\tan(bt_c) = \frac{\sin(bt_c)}{\cos(bt_c)} = \frac{-\sqrt{I_0}}{\left(\sqrt{S_0} - \frac{s}{b}\right)}$.

Thus, $\tan(\pi - bt_c) = \frac{\sqrt{I_0}}{\left(\sqrt{S_0} - \frac{s}{b}\right)}$ and $\pi - bt_c = \arctan\left[\frac{\sqrt{I_0}}{\left(\sqrt{S_0} - \frac{s}{b}\right)}\right]$, which yields

$$t_c = \frac{1}{b}\left(\pi - \arctan\left[\frac{\sqrt{I_0}}{\left(\sqrt{S_0} - \frac{s}{b}\right)}\right]\right) \tag{97}$$

Example 3

For our continuing model with $\beta = .03$, $r = .3$, $S_0 = 225$, $I_0 = 11$, $S^* = 100$, and $S_\infty = 16$, we have $t_c = 170.394$.

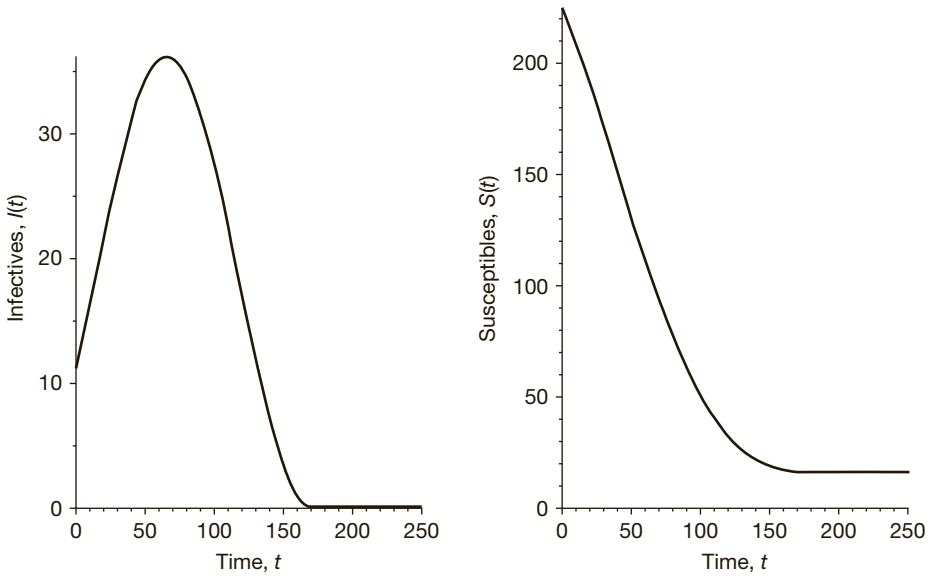


FIGURE 14.19 Results of the Mickens model showing the number of infectives and susceptibles over time.

To summarize: our explicit solutions to the Mickens model are

$$I(t) = \begin{cases} v(t)^2 & 0 \le t \le t_c \\ 0 & t > t_c \end{cases} \quad \text{and} \quad S(t) = \begin{cases} u(t)^2 & 0 \le t \le t_c \\ S_\infty & t > t_c \end{cases} \tag{98}$$

In Fig. 14.19, we show graphs of $I(t)$ and $S(t)$ as defined by Eq. (98) for typical values of the parameters.

We see that qualitative predictions of the Mickens model are similar to those of the classic Kermack-McKendrick SIR model while providing exact solutions of the differential equations. As Mickens observes the square root approach also provides two additional insights into the modeling process:

> *First, many sets of equations can provide the same qualitative features of a phenomena; second, in general, there are few a priori explicit rules that can be applied to restrict the structure of mathematical models. In other words, mathematical modeling is hard and thought must be put into deciding just what are the fundamental issues and concepts, and how they should be translated into the equations that will then be investigated.*

The Kermack-McKendrick and Mickens models both have the form

$$\frac{dS}{dt} = -\beta S^{\alpha} I^{\alpha}, \qquad \beta > 0$$

$$\frac{dI}{dt} = \beta S^{\alpha} I^{\alpha} - rI^{\alpha}, \quad r > 0$$

$$\frac{dR}{dt} = rI^{\alpha}$$

for some constant α. Mickens uses $\alpha = ½$ instead of $\alpha = 1$ as in the classic SIR model. I encourage you to experiment with other choices of α to see how the qualitative behavior of the dynamics of the model depends on α.

III. A Probabilistic Approach

A. A Stochastic Model of Simple Epidemics

In earlier chapters, there were several discussions of the weaknesses of the deterministic approach in models involving social or life sciences and the consequent need for probabilistic models. In addition to those general arguments, there are several that are particularly appropriate for models of epidemics. One of the fundamental assumptions of the deterministic models presented in Section II concerned the rate at which the susceptible population is reduced as an infectious disease spreads through a community. We assumed that the rate of change was proportional to the number of contacts between the susceptible and infective subpopulations and, further, that this number was proportional to the sizes of these two subgroups. In mathematical terms, we had the equation $dS/dt = -\beta S(t)I(t)$ in the classic SIR model and $dS/dt = -\beta\sqrt{S(t)}\sqrt{I(t)}$ in the Mickens approach. This axiom presumes that there is a homogeneous mixing of the members of the community. This is not a realistic presumption if the community is large enough to contain significant subgroups of different ages, interests, occupations, and geographical locations. The axiom is more likely to be accurate for a small group, such as recruits in an army camp or students living in a dormitory. But the smaller the group size, the less reasonable is a deterministic approach, since this approach is based on the hope that statistical fluctuations in behavior "smooth out" for large groups. These fluctuations are important in the study of small groups. The element of probability is of considerable importance here and should be incorporated into the mathematical model.

Even if our interest in the model is focused on the average number or expected value of susceptibles or infectives, a probabilistic approach is required. For, although with some population processes (such as the pure birth model of Chapter 10) the expected value was identical to the corresponding deterministic prediction, this is not always true of epidemic processes.

In this section we will develop a probabilistic version of the simple epidemic model of Section II.B. In the simple model, $R(t) = 0$ and $S(t) = N - I(t)$ for all t, so it is sufficient to investigate the function $I(t)$. The probabilistic model does not predict a precise value for $I(t)$ for each time t. Instead the model gives a set of probability distributions for $I(t)$—that

is, the model yields for each t and each nonnegative integer m, a probability—denoted $p_m(t)$—that there are exactly m infectives at time t:

$$p_m(t) = Pr(I(t) = m) \tag{99}$$

Note first that the number of infectives cannot exceed the total population N of the community at any time so that

$$p_m(t) = 0 \text{ for all } t \text{ if } m \geq N + 1 \tag{100}$$

The function $I(t)$ can then be thought of as a random variable that takes on possible values $0,1, 2, \ldots, N$. The simple epidemic model assumes that precisely one person in the community is an infective when the epidemic begins at time 0. Thus,

$$p_m(t) = \begin{cases} 1 & \text{if } m = 1 \\ 0 & \text{if } m \neq 1 \end{cases} \tag{101}$$

The probabilistic model of a simple epidemic is developed in the same spirit as the model for a pure-birth process presented in Chapter 10. The particular assumptions are these:

***Assumption* 1** A susceptible individual is infected when he comes into contact with an infective person.

***Assumption* 2** Once an individual is infected, he remains an infective for the remaining time.

***Assumption* 3** The probability that there is exactly one contact between a susceptible and an infective in a particular very short period of time is proportional to the number of susceptibles, the number of infectives, and the length of the time interval. In other terms, there is a positive constant β such that

$$\Pr(\text{Exactly one contact in time interval } (t, t + \Delta t) = \beta I(t)S(t)\Delta t = \beta I(t)(N - I(t))\Delta t$$

since there are no removeds in the simple model.

***Assumption* 4** The probability of more than one contact between susceptibles and infectives in a very short time period is negligibly small.

These assumptions are used to determine the probability of m infectives at time $t + \Delta t$, the number $p_m(t + \Delta t)$. There are three distinct and mutually exclusive ways that the community can have precisely m infectives at such a moment:

Event (A) There were exactly m infectives at time t and there was no contact between susceptibles and infectives during the period $(t, t + \Delta t)$.

Event (B) There were exactly $(m - 1)$ infectives at time t and there was precisely one susceptible-infective contact in $(t, t + \Delta t)$.

Event (C) There were less than $(m - 1)$ infectives at time t and there was more than one infective-susceptible contact in $(t, t + \Delta t)$.

By Assumption 4, the event (C) has negligible probability and can be ignored if Δt is very small. Thus,

$$p_m(t+\Delta t) = \Pr(\mathrm{A}) + \Pr(\mathrm{B}) \tag{102}$$

Now

$$\begin{aligned}\Pr(\mathrm{B}) &= \Pr(1 \text{ contact and } I(t) = m-1)\\ &= \Pr(1 \text{ contact} \mid I(t) = m-1)\ Pr(I(t) = m-1)\\ &= \beta(m-1)(N-(m-1))\Delta t p_{m-1}(t)\end{aligned}$$

by Assumption 3.

The probability of event (A) is computed similarly:

$$\begin{aligned}\Pr(\mathrm{A}) &= \Pr(0 \text{ contacts and } I(t) = m)\\ &= \Pr(0 \text{ contacts} \mid I(t) = m)\ Pr(I(t) = m)\\ &= (1 - \Pr(1 \text{ contact} \mid I(t) = m))\ Pr(I(t) = m)\\ &= (1 - \beta m(N-m)\ \Delta t)\ p_m(t).\end{aligned}$$

These equations yield the basic relationship

$$p_m(t+\Delta t) = (1-\beta m(N-m)\Delta t)p_m(t) + \beta(m-1)(N-m+1)\Delta t p_{m-1}(t) \tag{103}$$

which may be rewritten as

$$\frac{p_m(t+\Delta t) - p_m(t)}{\Delta t} = -\beta m(N-m)p_m(t) + \beta(m-1)(N-m+1)p_{m-1}(t) \tag{104}$$

Now let $\Delta t \to 0$ in Eq. (50) and obtain, as a limit, the differential equation

$$\frac{dp_m(t)}{dt} = -\beta m(N-m)p_m(t) + \beta(m-1)(N-m+1)p_{m-1}(t) \tag{105}$$

Our probabilistic model consists of this collection of differential equations, one each for $m = 1, 2, \ldots, N$ together with the initial conditions of Eq. (101).

B. Deductions from the Model

Examine the differential Eq. (105) first in the case $m = 1$. The equation takes the form

$$\frac{dp_1(t)}{dt} = -\beta 1(N-1)p_1(t) + \beta(1-1)(N-1+1)p_0(t) = -\beta(N-1)p_1(t) \tag{106}$$

which has the solution

$$p_1(t) = e^{-\beta(N-1)t} \tag{107}$$

The first prediction of this model is that there is always a positive probability that the epidemic does not spread beyond the original infective person—but that this probability decreases exponentially toward zero as time continues. This prediction follows from the fact that $\beta(N-1)$ is a positive constant and from knowledge of the exponential function $e^{-\beta(N-l)t}$.

Next consider the case $m=2$. The differential Eq. (105) then has the form

$$\begin{aligned}\frac{dp_2(t)}{dt} &= -\beta(2)(N-2)p_2(t)+\beta(2-1)(N-2+1)p_1(t)\\ &= -2\beta(N-2)p_2(t)+\beta(N-1)e^{-}\beta^{(N-1)t}\end{aligned} \tag{108}$$

so that

$$\frac{dp_2(t)}{dt}+2\beta(N-2)p_2(t)=\beta(N-1)e^{-\beta(N-1)t} \tag{109}$$

Eq. (109) is a first-order linear differential equation. It can be solved directly (see Appendix V) by making use of an integrating factor. In this case, the factor is $e^{2\beta(N-2)t}$ and the solution is found, using $p_2(0)=0$, to be

$$p_2(t)=\frac{N-1}{N-3}\left[1-e^{-\beta(N-3)t}\right]e^{-\beta(N-1)t} \tag{110}$$

This procedure can be continued to find $p_3(t)$ as the solution of the differential equation obtained by setting $m=3$ in Eq. (105) and making use of the explicit form of $p_2(t)$ together with the initial condition $p_3(0)=0$. The corresponding differential equation is

$$\frac{dp_3(t)}{dt} = -3\beta(N-3)p_3(t)+2\beta(N-2)\left[1-e^{-\beta(N-3)t}\right]e^{-\beta(N-1)t}$$

and the solution is

$$p_3(t)=\frac{(N-1)(N-2)}{(N-3)(N-4)(N-5)}\,\frac{(N-5)-2(N-4)e^{-\beta(N-3)t}+(N-3)e^{-2\beta(N-4)t}}{e^{\beta(N-1)t}} \tag{111}$$

In theory, we could continue in this fashion to obtain $p_4(t), p_5(t), \ldots, p_N(t)$ as explicit functions of t. The computations become formidable very quickly, however, and this is not an efficient procedure if N is moderately large. A more sophisticated mathematical technique permits a direct computation of the complex formula for $p_m(t)$ where m is any integer between 1 and N; the details can be found in Chapter 5 of Norman Bailey's *The Mathematical Theory of Epidemics* (see References).

C. A Comparison of Deterministic and Probabilistic Models

In the discussion of the pure-birth process in Chapter 10, we noted that the expected value of the population as given by the probabilistic model coincided exactly with the predicted value from the deterministic model. For models of epidemics, the expected values are, in

general, different from the solutions of the corresponding deterministic differential equations. We shall illustrate this for the models of a simple epidemic.

For each time t, the number of infectives is a random variable taking on values $1, 2, \ldots, N$ with probabilities given by $p_m(t)$, $m = 1, 2, \ldots, N$. The expected value of the number of infectives is

$$\varphi(t) = \sum_{m=1}^{N} m p_m(t) \tag{112}$$

The deterministic model predicts that the number of infectives at time t is given by the formula

$$I(t) = \frac{N}{1 + (N-1)e^{-\beta N t}} \tag{8}$$

In this section we will show that the two functions $\varphi(t)$ and $I(t)$ are not identical. There are a number of ways of doing this. One method would be to calculate $p_m(t)$ as an explicit function of t for each m and then use these to obtain an analytical description of $\varphi(t)$. As already noted, this approach leads to considerable computational difficulties. An alternative approach is to find some time t at which φ and I exhibit different properties. We will adopt this approach and show in particular that $\varphi''(0)$ and $I''(0)$ are different numbers.

Note first that the deterministic model gives

$$I'(t) = \beta I(t)(N - I(t))$$

so that

$$\begin{aligned} I''(t) &= \beta[I'(t)(N - I(t)) + I(t)(-I'(t))] \\ &= \beta I'(t)(N - 2I(t)) \end{aligned}$$

These equations give

$$\begin{aligned} I(0) &= 1 \\ I'(0) &= \beta(N-1) \end{aligned}$$

and

$$I''(0) = \beta\beta(N-1)(N-2) = \beta^2(N-1)(N-2)$$

Since $p_m(0) = 0$ if $m \neq 1$ and $p_1(0) = 1$, we have $\varphi(0) = 1$. Thus, $\varphi(0) = I(0)$. Examining the fundamental differential equation for the probabilistic model of a simple epidemic,

$$\frac{dp_m(t)}{dt} = -\beta m(N-m)p_m(t) + \beta(m-1)(N-m+1)p_{m-1}(t) \tag{105}$$

we find that

$$p_1'(0) = -\beta 1(N-1)p_1(0) + \beta(0)(N-1+1)p_0(0) = -\beta(N-1) \tag{113}$$

while

$$p_2'(0) = -\beta 2(N-2)p_2(0) + \beta(2-1)(N-2+1)p_1(0) = \beta(N-1) \tag{114}$$

and

$$p_m'(0) = 0 \text{ for all } m \geq 3 \tag{115}$$

Now we can calculate the derivative of the expected value:

$$\begin{aligned} \varphi'(t) = \frac{d\varphi}{dt} &= \frac{d}{dt}\sum_{m=1}^{N} m p_m(t) = \sum_{m=1}^{N} \frac{d}{dt}(m p_m(t)) = \sum_{m=1}^{N} m p'_m(t) \\ &= p'_1(t) + 2p'_2(t) + 3p'_3(t) + \cdots + Np'_N(t) \end{aligned} \tag{116}$$

Evaluation of the derivative at time $t = 0$ gives

$$\begin{aligned} \varphi'(0) &= p'_1(0) + 2p'_2(0) + 3p'_3(0) + \cdots + Np'_N(0) \\ &= -\beta(N-1) + 2\beta(N-1) + 3(0) + 4(0) + \cdots + N(0) \\ &= \beta(N-1) \end{aligned} \tag{117}$$

So far we have succeeded in showing that φ and I agree to the extent that $\varphi(0) = I(0)$ and $\varphi'(0) = I'(0)$. Finally, we evaluate $\varphi''(0)$. First, compute the expression for $\varphi''(t)$:

$$\varphi''(t) = \frac{d\phi'}{dt} = \sum_{m=1}^{N} m p''_m(t) \tag{118}$$

Now

$$\begin{aligned} \varphi''_m(t) &= \frac{d\varphi''_m}{dt} \\ &= \frac{d}{dt}[-\beta m(N-m)p_m(t) + \beta(m-1)(N-m+1)p_{m-1}(t)] \\ &= -\beta m(N-m)p'_m(t) + \beta(m-1)(N-m+1)p'_{m-1}(t) \end{aligned} \tag{119}$$

In evaluating $\varphi''(0)$ there will be precisely four nonzero terms in the sum of the right-hand side of Eq. (118). These will occur when the index m is 1, 2, and 3:

$$\begin{aligned} \varphi''(0) &= 1(-\beta)(1)(N-1)p'_1(0) + 2(-\beta)(2)(N-2)p'_2(0)) \\ &\quad + 2\beta(1)(N-2+1)p'_1(0) + 3(\beta)(3-1)(N-3+1)p'_2(0) \\ &= -\beta(N-1)(-\beta(N-1)) + 2[-2\beta(N-2)\beta(N-1)] \\ &\quad + \beta(N-1)(-\beta(N-1)) + 3[2\beta(N-2)\beta(N-1)] \\ &= \beta^2(N-1)(N-3) \end{aligned} \tag{120}$$

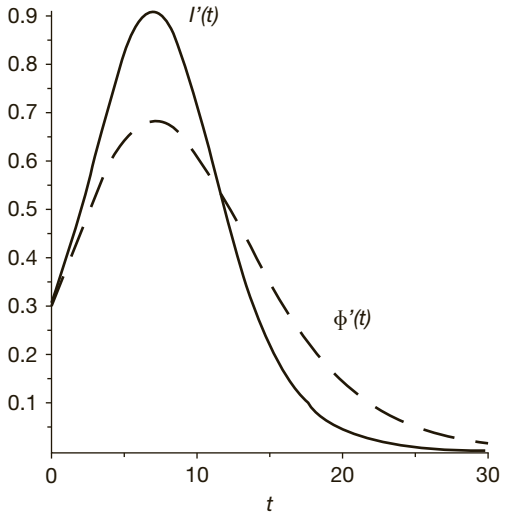


FIGURE 14.20 Comparison of deterministic (solid) and probabilistic (dashed line) epidemic curves for a simple epidemic with $S_o = 10$, $I_0 = 1$, and $\beta = .03$.

Thus, we have $I''(0) = \beta^2(N-1)(N-2)$ while $\varphi''(0) = \beta^2(N-1)(N-3)$ so that the functions $I(t)$ and $\varphi(t)$ are not identical.

Note that the ratio $I''(0)/\varphi''(0) = (N-2)/(N-3)$ is greater than one, so that the function $I'(t)$ is initially growing at a faster rate that $\varphi'(t)$. In fact, it can be shown that for all $t > 0$, $I(t)$ is always greater than $\varphi(t)$. Note also that for large values of N, the ratio is close to 1, so that the graphs of the functions $I(t)$ and $\varphi(t)$ as functions of t will be similar, at least close to $t = 0$. The graph of $\varphi(t)$ does not exhibit the same symmetry as the graph of $I(t)$, but the times at which both curves reach their maximum values are close together. The graphs of these functions for $N = 11$ are shown in Fig. 14.20.

The deterministic model of a simple epidemic offered a direct solution with no great mathematical difficulty. The probabilistic model of the same simple epidemic involves considerably greater complication. The development of a probabilistic model for the general epidemic as discussed in Section II.C is beyond the scope of the mathematics introduced in this text, as are models that are even more realistic in their assumptions.

IV. Historical and Biographical Notes

A. The Development of Mathematical Epidemiology

In an historical sketch of epidemics written in the early 19th century, a Vermont country doctor, Joseph A. Gallup [1815], noted that

> *Epidemic diseases, and their sequelae, occupy a very large portion of the catalogue of human maladies. No section of the globe is exempt from their ravages, and no society of individuals has been excused from weeds of mourning by the devastation of these scourges of man. They follow wherever the footsteps of man lead the way, and his traces are bestrewed with monumental inscriptions of human frailty.*

Records of epidemics appear in the literature of the ancient Greeks, and the Bible reports epidemics occurring among the Egyptians several thousand years earlier. Statistical information on the incidence and locality of cases of infectious diseases was first collected by two Englishmen, the statistician John Graunt (1620–1674) and the political economist and physician Sir William Petty (1623–1687) during the 17th century.

In 1840, William Farr (1807–1883), another English statistician and commissioner of the census, published some of his studies on statistical information, which he hoped would lead to the discovery of empirical laws on the growth and decline of epidemics. By a detailed examination of the spatial and temporal pattern of outbreaks of cholera, the English physician John Snow (1813–1858) demonstrated in 1855 that the disease was being spread by the contamination of water supplies. (Snow is also remembered as the man who introduced into English surgical practice the use of ether as an anesthetic.) In 1873, William Budd (1811–1880), established a similar manner for proving the spread of typhoid. Budd, another English physician, advocated disinfection as a method of preventing the spread of contagious diseases and recommended measures that stamped out Asiatic cholera and rinderpest in his country.

A coherent, predictive theory of epidemics requires both the development of sufficiently powerful mathematical techniques and the formation of sufficiently precise hypotheses about the spread of diseases that are suitable for expression in mathematical terms. The research achievements of biological scientists, especially those of Louis Pasteur and Robert Koch in the second half of the 19th century, established the physical bases for the cause of infectious disease and made possible both the mathematical modeling of epidemics and, more important, the public health measures that have lessened the chances of widespread epidemics.

Pasteur (1822–1895), the famous French chemist, is known for discovering that bacteria were the cause of anthrax, for developing successful vaccines against anthrax and cholera, and for pioneering treatment of hydrophobia in humans and rabies in dogs. Pasteur also isolated the bacilli causing two distinct diseases of silkworms and found a method for preventing the spread of these diseases, thereby saving the French silk industry.

A German physician and bacteriologist, Koch (1843–1910), was the first person to isolate and obtain a pure culture of the anthrax bacillus, to isolate tubercle bacillus, and to identify the comma bacillus as the cause of Asiatic cholera. He traveled to South America to study rinderpest, to India to study bubonic plague and cholera, and to Africa to learn more about malaria and sleeping sickness. Koch was awarded the 1905 Nobel Prize in physiology and medicine.

The first deterministic model of epidemics appeared in the English medical journal *Lancet* in 1906. This model, created by William H. Hamer, stressed the fundamental assumptions that the continuing spread of an epidemic depends on the number of susceptibles and the rate of contact between susceptibles and infectives. Hamer's mathematical assumptions, in one modified form or another, appear in almost all deterministic and probabilistic models of epidemics.

Beginning in 1911, Sir Ronald Ross published a series of books and papers developing a detailed deterministic model of malaria. A British physician born in India in 1857, Ross began his research into malaria in 1892 and after five years of patient work was able to piece together the life history of the malarial parasite in mosquitoes. Ross's work earned him the 1902 Nobel Prize for physiology and medicine. In addition to his medical and

mathematical writings, Ross published a novel, was a professor of tropical sanitation, and was director-in-chief of the Ross Institute and Hospital for Tropical Diseases in London. He died in 1932.

More elaborate deterministic models were developed by Kermack and McKendrick in a succession of papers published between 1927 and 1939. Perhaps their most important discovery was the Threshold Theorem discussed here in Section II.C. This result helped researchers account for the absence or occurrence of outbreaks of many epidemic diseases. Although McKendrick is best known for his work on deterministic models, he was also the first person to publish (in 1926) a probabilistic account of an epidemic process. Other pioneers in the use of probabilistic models were Major Greenwood in England and Lowell J. Reed and Wade Hampton Frost in the United States.

There were a number of important advances in mathematical epidemiology in the 1940s and 1950s, including Norman T. J. Bailey's complete solution of the probabilistic model for a simple epidemic (1950). In 1957, Bailey published the first textbook giving a systematic treatment of the whole field of mathematical modeling of epidemics. In a survey article 10 years later, Klaus Dietz noted that since Bailey's book appeared, "the contributions to this subject have themselves behaved like an epidemic." Bailey more recently wrote a subsequent edition, under the title *The Mathematical Theory of Infectious Diseases*, which provides a modern survey of this fast-growing discipline.

In recent years, mathematicians have developed and analyzed more complicated models of epidemics using new results about the qualitative behavior of nonlinear ordinary and partial differential equations and exploiting the power of digital computers to carry out simulations of systems involving many equations. Other researchers are using insights gained from the study of models of the spread of infectious diseases to study other dynamic systems. As we have seen, variations of the basic SIR model can be used to study the spread of rumors or drinking behavior. The epidemic models have also been starting points to examine eating disorders such as bulimia (González, 2003), fanatic behavior (Castillo-Chavez, 2003), and domestic violence (Abdul-Karim, 2012). Even zombies have fallen victim to mathematical modeling; the 2007 paper by Philip Munz et al. received much attention in the popular press.

B. Ronald Mickens

Physicist Ronald Elbert Mickens was born in the then deeply segregated city of Petersburg, Virginia, on February 7, 1943. His maternal grandfather introduced him to science as well as to Br'er Rabbit stories and folk medicine. By age 8, Mickens knew he wanted to be a scientist; he ultimately took enough extra summer courses to graduate early from Peabody High School (originally named "The Colored High School").

An avid reader as a youth, Mickens reports that he went to the public library daily. African Americans were restricted to the basement level of the building, which contained the card catalog of all the books in the building. The librarian often went upstairs to get books on science, rocketry, and calculus for Mickens. Eventually, Mickens joined a large group of fellow students who staged a successful demonstration that led to the library's racial integration.

Photo courtesy of Ron Mickens

Ronald E. Mickens

At age 17, Mickens entered Fisk University in Nashville, where he focused on mathematics, physics, and chemistry, ultimately graduating with a bachelor's degree in physics and one of the highest grade point averages in history of Fisk, along with Phi Beta Kappa honors. A recipient of fellowships from the Woodrow Wilson and Danforth foundations, Mickens received his doctoral degree in physics from Vanderbilt University. A National Science Foundation Postdoctoral Fellowship gave him an opportunity for further study and research before he returned to Fisk to teach physics. In 1982, Mickens moved to Clark Atlanta University, where he currently holds the title of Distinguished Fuller E. Callaway Professor of Physics.

Mickens has conducted research in the areas of complex functions, theoretical elementary particle physics, mathematical epidemiology, and modeling of nonlinear oscillations. In addition to scores of journal articles, he has written numerous advanced texts, including *Mathematics and Science*, *Difference Equations*, *Applications of Nonstandard Finite Difference Schemes*, *Mathematical Methods for the Natural and Engineering Sciences*, *An Introduction to Nonlinear Oscillations*, *Oscillations in Planar Dynamic Systems*, and *Truly Nonlinear Oscillations: Harmonic Balance, Parameter Expansions, Iteration, and Averaging Methods*.

In addition to research and teaching, Mickens has devoted significant efforts to open physics to minority students. He serves as historian for the National Society of Black Physicists. Among his other books are a history, *The African American Presence in Physics*, and a biography, *Edward Bouchet, The First African-American Doctorate*. Both Fisk and Vanderbilt are located in Nashville, which served as the base of the Student Nonviolent Coordinating Committee (SNCC), one of the most active organizations in the civil rights movement of the 1960s. Mickens participated in a number of SNCC-sponsored events.

You can learn much more about Mickens's life and work from an extended (163-minute) video interview with him available at http://www.idvl.org/sciencemakers/Bio13.html. At its conclusion, Mickens observes,

> *I'm having fun, and I've had fun for a long time. And I wish that there were more people who could see intellectual activities as things that can provide them with great joy, with great fun, and that it also provides you the opportunity to meet all kinds of interesting people, to go to all kinds of strange and interesting and weird places.*

EXERCISES

DETERMINISTIC MODEL OF A SIMPLE EPIDEMIC

1. Use the relation $dI/dt = \beta I(N - I)$ to sketch a graph of dI/dt as a function of I. Can you draw any conclusions about the model from this graph?

2. Derive Eq. (8) from Eq. (7).

3. Show that Eq. (6) gives $I''(t) = \beta^2 I(N - I)(N - 2I)$. Does this result imply that the maximum value of $I'(t)$ necessarily occurs when $I = N/2$?

4. Show that Eq. (11) can be derived from Eq. (8) and the relation $dI/dt = \beta I(N - I)$ without further differentiation.

5. Discuss how you would obtain a numerical value for β from observed information concerning the number of infectives and susceptibles at various times.

6. Show that $I'(t_{max} + a) = I'(t_{max} - a)$ for all α—that is, verify the claim that the epidemic curve is symmetric about the vertical line through t_{max}.

7. Show that $S(t)$ drops below 1 as soon as t exceeds $2t_{max}$. In what sense does this observation justify the claim that "the simple epidemic is over by time $2t_{max}$";?

8. Explain why $\lim_{t\to\infty} \frac{N-1}{e^{\beta N t}} - 1 = -1$

9. Generalize the simple model by allowing β to be a continuous function, $\beta(t)$, of time t, rather than a simple constant. In particular, discuss the consequences of the model in each of the following cases:

(a) $\beta(t)$ is an increasing function—for example, $\beta(t) = t$.

(b) $\beta(t)$ is a decreasing function—for example, $\beta(t) = e^{-t}$.

(c) $\beta(t)$ is a periodic function—for example, $\beta(t) = \cos t$.

10. Generalize the simple model to situations in which the size of the total community changes during the course of the epidemic because of births or from deaths due to causes other than the infectious disease. In particular, investigate the simple model if $N(t)$ is growing

(a) exponentially

(b) logistically

11. Investigate the simple model if $1(0) = I_0$ is greater than 1. Find analogues of Eqs. (8) and (11) in particular.

12. Sketch the epidemic curve if $N = 1000$, $1_0 = 100$, and $\beta = .003$. Use the results of Exercise 11.

13. When the early stages of an epidemic are observed, steps are often taken to prevent its spread. Suppose, for example, that public health officials administer vaccines at a constant rate of α inoculations per time unit. Suppose this program continues until the entire population is either vaccinated or infected. The mathematical model in this situation might take the form

$$dI/dt = \beta I(t)(N - \alpha t - I(t))$$

since $N - \alpha t - I(t)$ represents the number of susceptibles.

(a) Why is this a reasonable model?

(b) Show that an epidemic described by this model ends when $t = N/\alpha$.

(c) What predictions can you make from this model without solving the differential equation explicitly?

(d) Can you solve the differential equation?

Discrete Version of the Simple Model

14. Show that the maximum value of $f(x) = x + \beta x(N - x)$ does occur at the average of $1/\beta$ and N. Determine that maximum value.

15. Let $P_k = I_k/N$.

(a) Show that the equation $I_{k+1} = I_k + \beta I_k(N - \mathrm{I}_k)$ becomes $P_{k+1} = P_k + \beta P_k(1 - P_k)$.

(b) Show that the equation in (a) is the discrete logistic growth model. Is there a choice of β that leads to chaos? Recall our discussion in Chapter 3, but note that $0 < \beta < 1$.

16. Formulate and analyze the discrete version of the model introduced in Exercise 12.

DETERMINISTIC MODEL OF A GENERAL EPIDEMIC

17. A simpler model than the one discussed in the text is based on the equations

$$\frac{dS}{dt} = -\beta S_0 I$$

$$\frac{dI}{dt} = \beta S_0 I - rI$$

$$\frac{dR}{dt} = rI$$

where β and r are again positive constants measuring the rates at which susceptibles become infected and infective individuals are removed.

(a) Determine S, I, and R as explicit functions of t if initial numbers are S_0, I_0, and $R_0 = N - S_0 - I_0$.

(b) Prove that if $\beta S_0 < r$, the disease will not produce an epidemic.

(c) Discuss what happens in the case $\beta S_0 > r$.

18. In discussing the limiting behavior of R, S, and I we made use of a theorem stating that if $f(t) \le N$ for all t and if f is monotonically nondecreasing, then there exists a number $L \le N$ such that $\lim_{t\to\infty} f(t) = L$. Find a proof of this theorem.

19. Use Eq. (27) to sketch a graph showing the relation between S and R.

20. Apply the bisection technique to the function $f(x) = x^2 - 2$ to find an approximation of $\sqrt{2}$ accurate to two decimal places.

21. Apply the bisection technique to Eq. (39) with $S_0 = 1000$, $N = 1001$, $r = .9$, and $\beta = .002$ to estimate S_∞, to the nearest integer.

22. Newton's method is a technique for finding roots of the equation $f(x) = 0$ when f is a differentiable function. It often is a more efficient technique than the bisection method. Most calculus texts will contain some discussion of Newton's method. Investigate how you might apply Newton's method to calculate S_∞.

23. Eq. (34) defines I as a continuous function of S:

$$I = g(S) = N - S + p\log\left(\frac{S}{S_0}\right)$$

Show that $g(S_0) > 0$ and that $g\left(\frac{S_0}{e^{-N/p}}\right) < 0$.

24. Find the exact solution to Eq. (41). See Bailey's text if you get stuck.

25. Solve Eq. (41) if a Taylor series approximation for e^x is used when the series is terminated after

(a) One term

(b) Two terms

(c) Four terms

26. Show that the epidemic curve is not symmetric about any vertical line.

27. Use Eqs. (43) and (44) and the approximation $R_\infty \sim 2v$ to obtain various estimates for R_∞ in the case $\beta = .001$, $r = .9$, $N = 1001$, and $I_0 = 1$. Compare results with the number obtained by using the bisection technique.

28. Repeat Exercise 27 using the data from Exercise 21.

DISCRETE VERSION OF THE MORE GENERAL EPIDEMIC MODEL

29. Implement the first discrete version of the Kermack-McKendrick model. Find conditions on the parameters under which the entire susceptible population becomes infected in a finite time.

30. In comparing his model with data collected over a 15-year period, Spicer found that the model yielded a higher number of infectives than seen in the data. He concluded that in the early stages of an epidemic, the actual number of cases is underreported. Is that a reasonable claim? Explain.

31. Implement the Spicer model with $p_j = (0.9)^j$.

(a) How does the graph of the number of infectives compare with Fig. 14.9?

(b) Examine the graph of infectives if $\beta = .003$ and $\beta = .3$.

(c) How many days does it take for the number of infectives to drop below 1 for different choices of β?

RUMORS

32. Why is the number of distinct contacts between a pair of yentas equal to $\frac{y(y-1)}{2}$?

33. For our rumor model, show that $dz/dt = y(y-1+z)$.

34. Carry out the details of solving $dy/dx = -2 + N/x$ with initial conditions $y = 1$ when $x = N$ to obtain the result stated in Eq. (52).

35. Let $f(U) = 2(1-U) + \ln(U)$ for $U > 0$.

(a) Show that $f(1) = 0$.

(b) By examining the first and second derivatives of f, graph the function on the interval (0, 2).

(c) Prove that there is exactly one other positive number $U^* \neq 1$ for which $f(U^*) = 0$.

(d) Use the bisection technique or other numerical algorithm to approximate U^* to three decimal places.

36. Suppose not every conversation between a gossiper and another individual contains the rumor, but only in a certain percentage α of the encounters does the gossiper attempt to pass on the rumor. How does this assumption change the equations of our model? How does it affect the long-term spread of the rumor?

37. A yenta may become a stifler either by vowing never to pass on a rumor or simply by forgetting the rumor. Modify our model to add this feature and analyze the resulting dynamics of the system.

38. (Adapted from Daley and Gani.) A "stifling experience" occurs when a yenta meets another yenta or a zapper. In our simple rumor model, we assumed that a single stifling experience was enough to convert a yenta into a zapper. In the "k-fold stifling model," we assume that a yenta is not converted until the yenta has had k stifling experiences. Let y_i (for $i = 1,2,\ldots,k$) be the number of yentas who have had $i-1$ stifling experiences.

(a) Show that $y = y_1 + y_2 + \cdots + y_k$.

(b) Show that the k-fold stifling model may be represented by the system of differential equations.

$$
\begin{aligned}
dx/dt &= -xy \\
dy_1/dt &= xy - y_1(y-1+z) \\
dy_i/dt &= (y_{i-1} - y_i)(y-1+z), (i = 2,\ldots,k) \\
dz/dt &= y_k(y-1+z)
\end{aligned}
$$

(c) Show that the functions $x(t)$, $y_i(t)$ $(i = 1,\ldots,k)$ satisfy

$$
\begin{aligned}
&(k+1)x(t) + ky_1(t) + \ldots + y_k(t) + N\ln(N/x(t)) \\
&\quad = (k+1)N + k \text{ for all } t \geq 0
\end{aligned}
$$

(d) Show that the analogue of U in Eq. (54) is the root in (0, 1) of $(k+1)(1-U) + \ln U = 0$.

URBAN LEGENDS

39. Set up the Noymer model and replicate our findings with software that will generate graphical solutions to systems of differential equations.

40. The graph in Fig. 14.10 indicates that the first outbreak of the urban legend rumor is the most intense. In subsequent outbreaks, the maximum number of infected individuals is smaller, but the time durations of the outbreaks appear longer. Can you provide a verbal explanation for these properties?

41. How far apart are the peaks of the outbreaks? Why?

42. Run the Noymer model with $\upsilon = .2$ and $\lambda = .0012$.

43. How would you modify the Noymer model to keep track the total number of people who ever believed the urban legend?

44. Our urban legend model assumed the same recovery rate held for all contagious individuals. How might the model be modified to account for "diehards," individuals Noymer describes as those "who won't let the story rest and who find the occasional recruit, even after a recent epidemic has wiped out most susceptibles in the population." How does the existence of diehards affect the long-term persistence of the urban legend?

PROBLEM DRINKING

45. Show that the epidemic will not necessarily die out if $\mathfrak{R}_\phi < 1$, but if $\mathfrak{R}_\phi > 1$, there will be a long-term persistent population of problem drinkers.

46. Let $f(x) = x^2 - Bx + C$ and suppose that B and C are positive, $f'(1) > 0$, and $B^2 - 4C > 0$.

(a) Show that $f'(1) > 0$ implies that $B < 2$ and hence $B/2 < 1$.

(b) Show that $f(x) = 0$ has two distinct roots x_1 and x_2. Let x_1 be the smaller root.

(c) Show that $C > 0$ implies that $0 < x_1 < B/2 < 1$.

(d) Show that the minimum value of f is negative and occurs at $x = B/2$.

(e) Show if that if $B < 1$, then x_2 is also less than 1.

MICKENS MODEL

47. Show that equations of the Mickens model imply that the total population remains constant. [Hint: Add them up.]

48. Use the fact that $\frac{dI}{dt}\big|_{t=0} = \sqrt{I_0}\left(\beta\sqrt{S_0} - r\right)$ to show that if an epidemic begins at time $t = 0$ with a positive number of infectives, then it can spread if and only if $S_0 > \left(\frac{r}{\beta}\right)^2$.

49. Show that $R_0 = \left(\frac{\beta}{r}\right)^2 S_0$.

50. Suppose there are no susceptibles in the population but that there are some number of infectives.

(a) Show that in the Kermack-McKendrick model, $I(t) = I_0 e^{-rt}$, so that there will always be some number of infectives. Why is this an unrealistic prediction?

(b) In contrast to (a), show that in the Mickens model, the number of infectives becomes 0 at a finite time $t^* = \frac{2\sqrt{I_0}}{r}$. [Hint: Solve the differential equation $\frac{dI}{dt} = -r\sqrt{I}$.]

51. Carry out the details to show that $\frac{dv}{dt} = bu - s$.

52. Carry out the details to show that $v'' = -b^2 v$.

53. Verify that the formulas for the initial conditions are correct.

54. Verify that the proposed solutions for u and v are correct by substituting the formulas for $u(t)$ and $v(t)$ into the second-order differential equations.

55. Use the techniques in Appendix V (or from a Linear Algebra or Differential Equations course) to derive the solutions of the Mickens model.

56. Find an explicit formula for $R(t)$ in the Mickens model.

57. Determine the time t_m when the infective population reaches its maximum value I_{max}.

58. Find the time when the disease is spreading most rapidly—that is, when dI/dt reaches its maximum value.

59. Prove that the functions $I(t)$ and $S(t)$ as defined by Eq. (98) are both continuous and differentiable at $t = t_c$.

60. Suppose that the susceptibles, infectives, and removed have a constant per capita death rate independent of the diseases and that new susceptibles enter the population at a constant birth rate.

(a) Show that a reasonable model for an epidemic under these conditions has the form

$$\frac{dS}{dt} = \lambda - \mu S - \beta S^\alpha I^\alpha$$

$$\frac{dI}{dt} = \beta S^\alpha I^\alpha - (\mu + r)I$$

$$\frac{dR}{dt} = rI - \mu R$$

(b) If $N(t) = S(t) + I(t) + R(t)$ represents the total population at time t, show that the model in (a) implies that $\frac{dN}{dt} = \lambda - \mu N$.

(c) If the initial total population is $N(0) = N_0$, use the differential equation in (b) to show that $N(t) = \frac{\lambda}{\mu} + \left(N_0 - \frac{\lambda}{\mu}\right)e^{-\mu t}$. What happens to the population in the long term?

PROBABILISTIC MODEL FOR SIMPLE EPIDEMIC

61. Graph $p_2(t)$ as a function of t. For what value of t does it reach a maximum?

62. **(a)** Carry out the details of determining $p_3(t)$ and graph the function.

(b) Determine $p_4(t)$ and sketch its graph.

63. Show that Eq. (105) is valid for $m = 1$, although the derivation given for Eq. (105) does not quite work.

64. Prove that $p_0(t) = 0$ for all t for the given model.

65. Work out some details of the model if the initial population of infectives, I_0, is greater than 1.

66. Can $\varphi(t)$ be computed using the technique of Exercise 51 of Chapter 10?

67. What conclusions can you make about the variance of the number of infectives for the probabilistic model with $I_0 = 1$?

68. Investigate Bailey's explicit solution of the probabilistic model.

SUGGESTED PROJECTS

1. Modify the deterministic model to allow for a constant, nonzero infectious period—that is, assume that a constant number of days must pass between the time a person becomes infective and the time when he can transmit the disease to others. Analyze such a model and interpret the mathematical conclusions.

2. Generalize the deterministic model of Section II.C to allow for a nonconstant population, $N(t)$. In particular, investigate the model if the population is growing exponentially or logistically.

3. Formulate a probabilistic version of the deterministic model of Section II.C and analyze it in the spirit of Section III.

4. Soper [1929] proposed a mathematical model for the spread of measles that he believed adequately explained the recurrence of measles epidemics. Study his model and some of the corrections and extensions that have been made to it.

5. Investigate the De Hoog et al. [1979] discrete version of the Kermack-McKendrick model. Derive the analogue to the threshold theorem in the continuous case they found. What features do the continuous and discrete versions share? How do they differ in the predictions?

6. The myxoma virus causes the disease Myxomatosis in rabbits. The virus was deliberately introduced into the United Kingdom and Australia in an effort control rabbit infestation and population there. A bacteriologist accidentally introduced the disease in France when he used the virus to rid his private estate of rabbits in June 1952. Within two years, 90% of the wild rabbits in France were dead. Examine Ian W. Saunders's discrete version of the general epidemic model [1980], which he developed to study the dynamics of the disease in Australia. In his model, Saunders incorporates a latency period of seven days during which the rabbit is infected but cannot spread the disease.

7. Investigate other variations of the Mickens model in which you replace $S^{\alpha}I^{\alpha}$ by a suitable function $f(S, I)$ where $f(0, I) = 0$ and $f(S, 0) = 0$.

8. Mickens [2010] explores the model of Exercise 15 for $a = ½$. Investigate his treatment of both the original system of differential equations and their discrete analogues.

9. Formulate and analyze a probabilistic model for the spread of rumors. See the discussion in Chapter 5 of Daley and Gani for suggested approaches.

10. Investigate the application of epidemic models to eating disorders. Start by working through the paper by González et al. [2003].

You can find a listing of references and suggestions for additional reading on the book's website, www.wiley.com/college/olinick

CHAPTER 15

Roulette Wheels and Hospital Beds: A Computer Simulation of Operating and Recovery Room Usage

We are more than half what we are by imitation. The great point is to choose good models and to study them with care.

—Philip Dormer Stanhope, *Earl of Chesterfield*

I. Introduction

A. The Need for Simulation

In previous chapters we have seen that we can successfully attack a wide variety of problems by modeling their essential features with mathematical concepts and then using the analytical tools of the mathematician to make predictions about a system's behavior. There are many problems in the social and physical sciences, however, that do not appear to be amenable to solution by currently available analytic methods.

The mathematical modeling approach can break down in two essentially different ways. If we re-examine the basic diagram for model building (Fig. 15.1), we note where the difficulties may arise.

In the first place, we must translate the important features of the real-world phenomenon into mathematics. But which branch of mathematics do we choose? For some problems, there seem to exist several different classes of techniques from which we can choose. A deterministic approach using differential equations may suggest itself, or perhaps a probabilistic scheme using Markov chains. The history of scientific thought reveals many instances when a branch of pure mathematics was seized upon as the proper vehicle for a study of real-world phenomena. To develop his theory of general relativity, for example, Albert Einstein made use of non-Euclidean geometries, a subject previously considered by many to be frivolous and entirely lacking in applicable content. For some real-world systems, however, the complexity and variety of the interactions among the important variables—as we understand them—do not seem to fit any existing part of mathematics. In such an instance, the modeler may have to create new mathematical tools. There is much evidence that many important parts of mathematics were developed to

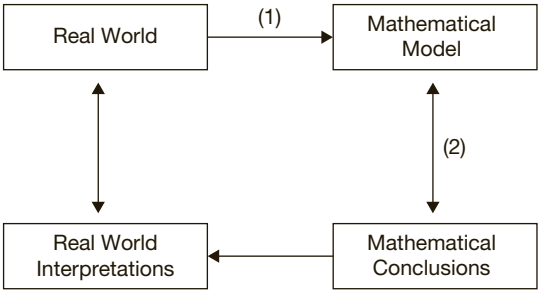


FIGURE 15.1 The schematic diagram illustrating the modeling process.

provide models for problems in the physical sciences; the same process is occurring in the social sciences as well. See Chapter 16 on Game Theory for such an example.

A second kind of difficulty occurs in taking the step from mathematical model to mathematical conclusions. The modeler may believe, for example, that the problem she is interested in can best be formulated as a question in geometric topology. It may turn out, however, that the topological question has not yet been settled, but is an unsolved research problem in mathematics. The work of Isaac Newton provides an example of this sort of difficulty. In the 17th century, Newton used his celebrated laws of motion and gravitational attraction to formulate a mathematical model for the relative motion of two bodies. The model was a differential equation requiring 12 integrations to solve explicitly. Newton worked these out and, taking the bodies to be the sun and a planet in the solar system, he showed that the model's predicted behavior of planetary orbits was in precise agreement with the laws of planetary motion that had been empirically determined by Johann Kepler. This achievement was one of the great milestones in the history of thought. When Newton turned to the analysis of the interactions of three bodies (for example, the interaction of gravitational attractions of the earth, moon, and sun), his laws led to another differential equation. An explicit solution for this equation required 18 successive integrations. Newton was unable to carry out the integrations completely. Neither was any other mathematician, physicist, or astronomer during the next 200 years. Finally, in the late 19th century, Henri Poincaré showed that it was impossible to get an exact solution for Newton's equation; further progress on the problem could only be made by approximation techniques.

Thus, the mathematical modeler may have to wait until new branches of mathematics are created or unsolved problems in existing branches are resolved before she can obtain a valid and useful model. In many real-world situations, however, there is a demand that a "solution" to the problem be found immediately.

In such situations, the modeler will often abandon theoretical formulations for simulation models. Simulation is a dynamic act of imitation of one or more essential features of a system. In a simulation we try to copy the behavior of a process where the possible causes and outcomes are fairly well understood while the relationships among them may be quite complex and incapable of simple analytic description. In this chapter, we will focus on *discrete-event simulation*. Chapter 19 provides an introduction to *agent-based simulation*.

It is easier to grasp the concept of simulation by examining several examples than by attempting to present additional definitions.

B. Examples of Simulation

A company that had been manufacturing elevators for many years decided to produce electric buses for public transportation. After some preliminary testing, the company built a prototype electric bus and offered it to the city of San Francisco for a free trial demonstration. Mass transit officials in the city wanted to know how well the electric bus would perform on a heavily traveled route that climbs a steep hill. They discussed replacing the gasoline-powered bus that normally serviced that route with the electric bus for a day so that they might assess how well the new bus would climb or fail to climb the hill when it was filled with passengers.

The city's safety officer objected that if the electric bus failed and slid down the hill out of control, many people could be injured. He suggested that the transit officials simulate an actual ascent of the hill by loading the electric bus with sandbags whose total weight and distribution inside the bus would resemble a busload of people. A test run could then be made with the sandbags. If the bus failed, the social costs of the accident would then be far less than if the bus were crammed with people.

The suggested test run is a *simulation* of an actual one. Most of the important features of an actual run are present in the simulated one. If the bus loaded with sandbags should fail, then it is likely that the same bus with an equal weight of humans aboard would also fail. But what if the bus succeeds in making it to the top of the hill and back down again without incident during the simulation? Is this a guarantee that it will do as well with living cargo? After all, the people may move around inside the bus while it is moving and may even take it into their heads to begin rocking the vehicle. The sandbags cannot imitate this behavior. Even if we feel that the motion of the passengers is not a critical factor, there is still a question of how many simulated runs we should try before declaring the bus safe. We shall return to these kinds of objections later in the chapter.

Such a simulation, intended to provide a safe way to test a new form of transportation, may itself contain unexpected dangers. Simulations can end in tragedy. Here is an excerpt from a *New York Times* article describing one:

> ***TRAIN TO KENNEDY DERAILS IN A TEST***
>
> *A futuristic three-car elevated train, the precursor of a $1.9 billion automated light-rail system that is expected to carry millions of air travelers a year to and from Kennedy International Airport, derailed on a curve during a test run to the terminals yesterday, killing its operator, who was alone on board.*
>
> *Its speed unknown, the sleek white AirTrain . . . slammed into a concrete retaining wall 25 feet above ground. . . . The force gashed open the front car, which sheared away 150 feet of the wall and came to a halt with its right side partly overhanging the parapet.*
>
> *The cause of the crash was not immediately determined, but . . . investigators were looking into the possibility that 16,000 pounds of concrete ballast—put aboard to simulate a load of passengers—had shifted on the gentle curve, leading the front end, and then all three cars, to stray and jump the tracks . . .*
>
> *Moreover, under the force of the collision, investigators said, tons of the ballast in the front car slid forward, pinning and fatally injuring the train's operator.*

As a second example of simulation, consider the engineer who first proposed designing a giant jet aircraft with the engines mounted on the tail of the plane instead of

Photo by Richard Lee for *The New York Times*. Reproduced by permission of Richard Lee

FIGURE 15.2 Derailed AirTrain.

under the front wings. The cost of building a prototype plane of the actual size to be marketed would cost millions of dollars, all of which would be wasted if the plane couldn't get off the ground or crashed shortly after takeoff on a test flight. The engineer has the problem of testing her design without actually building the plane.

A possible solution for her is to construct a small, scaled-down version of the plane and to test this model in a wind tunnel. The stability of the model aircraft can be tested in a variety of wind conditions approximating those the full-sized plane might encounter in the air. If the model plane cracks up every time it is subjected to the equivalent of a 25-knot wind, perhaps there is something fundamentally wrong in its design. On the other hand, if the model performs well in the wind-tunnel experiments, the engineer can have more faith that her idea is a good one. In the experiment, the model plane simulates a real plane and the wind tunnel simulates actual wind conditions. The experiment has the advantage of being relatively inexpensive to perform, but it has drawbacks too. The tunnel may not provide a sufficiently realistic imitation of atmospheric conditions. The model plane is, of course, of a different size and made of different materials than the plane it is meant to simulate. The full-sized plane may not necessarily behave in the same fashion as its scaled-down version. Still, simulation may be the only reasonable, safe, quick, and inexpensive way to test the plane's design.

In this chapter, we will examine, in some detail, an example of a computer simulation. By now you are familiar with the use of the computer as a tool in the analysis of a theoretical model, principally as a source of high-speed numerical computation. In simulation experiments, the computer is employed as a substitute for a theoretical model. We define a set of numerical-valued variables to represent the principal features of the system to be simulated. Then we formulate a computer program as a set of instructions, representing the decision rules or laws that determine how the system's features are to be modified as time goes on. In principle, this sort of computation can be carried out by hand, using pencil and paper, but the computer gives us the enormous advantages of speed, tirelessness, and memory of intermediate results. The details of a particular computer simulation will be presented in Part V. First, we need to explain the problems we hope to solve by a simulation.

II. The Problems of Interest

Diagnosis and treatment of many medical disorders requires that a patient be hospitalized for one or more days. One of the major constraints on the service a hospital can render to the surrounding community is thus the number of hospital beds it has. It often happens that a person who has a nonemergency medical problem may have to wait several days to be admitted to the hospital because all its beds are filled.

When this kind of problem begins to affect too many patients of too many of the doctors on its staff, the hospital administration may plan to construct new hospital facilities to provide an increase in the number of its beds. If the hospital decides to expand its bed complement, then it must also consider the increased demands that will be made on other aspects of its operations. It must determine whether its current medical and nonmedical facilities are adequate to provide for the larger number of patients who will be in the hospital each day. The hospital must decide, for example, how many new nurses to hire, how much new equipment to order, whether to expand the pathology and pharmacy departments, and so on.

In this chapter, we will examine how simulation techniques can help assess the increased need for operating-room (OR) and recovery-room (RR) facilities that an expanded bed complement will produce. This was the problem studied by Homer H. Schmitz of the Deaconess Hospital in St. Louis, Missouri, and N. K. Kwak of St. Louis University. Deaconess Hospital was planning to add 144 medical-surgical beds to its currently existing facilities in the early 1970s. Schmitz and Kwak formulated three primary questions:

1. How many more surgical procedures will Deaconess Hospital perform because of the increased bed capacity?
2. How much operating room time and space will the surgical procedures require?
3. How much recovery room time and space will the surgical procedures require?

The first question was answered by a relatively simple extrapolation technique, while insight into the other problems was gained through a computer simulation.

III. Projecting the Number of Surgical Procedures

Schmitz and Kwak [1972] began their study by collecting information on hospital procedures in effect during the period when the expanded bed complement was still in its planning stages. An analysis of the hospital's records for 1970 indicated that 42% of medical-surgical (MIS) patients actually had surgery. Assuming that the relative proportions of medical and surgical patients would not be affected by an increase in bed complement, it is a simple matter to project that if 144 MIS beds are added, then approximately 60 of them (144×0.42) will be utilized by patients who have surgery.

A critical factor in estimating the number of surgical procedures that would be performed is the length of stay in the hospital for each patient. This, of course, would depend on the nature of the surgery. For example, before the expansion, 4.5% of the total number of surgical procedures performed at the Deaconess Hospital were ophthalmology cases.

Table 15.1 Increase in surgical cases based on increased bed count

Type of surgery	Increase in number of cases per year
Ophthalmology	132
Gynecology	282
Urology	264
Orthopedic	202
Ear-nose-throat (ENT)	1098
Dental surgery	715
Other major surgery	683
Total projected increase	3376

The average length of stay for these cases was 7.4 days. Of the 60 new beds that will be used by surgical patients, we can estimate that about $.045 \times 60 = 2.7$ beds will be used by ophthalmology patients. Imagine that a particular bed is set aside for ophthalmology patients. An average patient will occupy that bed for 7.4 days. Thus, during the year 49 patients $(365/7.4 = 49)$ will be able to be treated for each ophthalmology bed. Since there will be 2.7 new ophthalmology beds, during the year there will be 132 (2.7×49) new ophthalmology surgical procedures.

Using the assumptions of full-bed utilization and the same patient mix in the future as was experienced in the past, Schmitz and Kwak estimated the increases in surgical procedures for six other major types of surgery in the same way as for the ophthalmology cases. Their results are presented in Table 15.1.

These extrapolations were based on an actual count of 6,293 surgical procedures performed in 1970. Adding the projected total of 3,376 new procedures, we arrive at an estimate of 9,669 projected surgical procedures when the new bed complement is fully utilized.

The daily surgical load is then determined by dividing the annual number of surgical procedures by the number of days in the year. Thus, the hospital can expect to have 9669/365 surgical procedures on an average day. For the simulation procedure, this number is rounded to the nearest integer, 27.

IV. Estimating Operating Room Demands

A. Length of Stay in Operating Room

How many operating rooms will be necessary to perform 27 surgical procedures each day? This will depend on several factors, such as the time of day of the first operation, the length of time necessary to prepare an operating room for a new patient after an old one has left, and the number of hours per day that surgeons are willing to work. A principal factor will be the length of time each operation requires.

Let us illustrate with some crude estimates. Suppose each operation (including make-ready time for the next patient) takes exactly 1 hour. Suppose also that the surgeons will work only during the period between 8 o'clock in the morning and 5 o'clock in the afternoon. Then each operating room can be used for 9 hours each day, so 9 surgical

procedures can be completed in each room. It would then take 3 operating rooms to accommodate the expected 27 daily procedures.

Of course, in actual practice not every operation takes 1 hour. A tonsillectomy requires much less time than a heart transplant, for example. We might argue, however, that our crude estimate of three operating rooms would hold up if the *average* length of an operation is 1 hour. As we saw in our study of expected value in Chapter 10, there are difficulties that arise if we look only at average values. There can be quite a variation in the lengths of operations that still produce an average of 1 hour. Also, although the average length for the 9,669 procedures may well be close to 1 hour, this may not be the case for a randomly selected group of 27 of them.

Why do we say "randomly selected"? Although the hospital can estimate fairly accurately the total number of different types of surgical procedures to be performed over a 12-month period, it cannot predict the order in which the patients will present themselves for treatment. On some days, there may be a relatively large number of patients who need operations that will last more than 2 hours, while on other days almost all the procedures will be relatively minor ones. The hospital must be ready to handle these deviations from the average.

It is necessary, then, to examine more carefully the lengths of stays in the operating room. Schmitz and Kwak did this by collecting a sample of 445 surgical patients treated in 1970 at Deaconess Hospital. Data was collected on the type of surgery performed, the length of time spent in the operating room, and the number of days the patient was hospitalized. The percentages of the various types of surgery and the average length of stay for the total population of patients in the hospital were inferred from this sample data; in Section III, we saw how these numbers were used.

In Table 15.2, the actual and relative frequencies for length of stay in the operating room for the 445 patients of the sample are presented.

In Table 15.2 and in all subsequent discussion, time segments are given in hundredths of an hour rather than in minutes, because this simplifies the mathematical calculations. From Table 15.2, for example, we note that 2.9% of all operations lasted between 2.5 hours and 3 hours.

Table 15.2 Length of stay in the operating room

Length of stay in hours	Frequency	Relative frequency
0.01–0.50	181	40.7
0.51–1.00	103	23.2
1.01–1.50	64	14.4
1.51–2.00	42	9.4
2.01–2.50	22	4.9
2.51–3.00	13	2.9
3.01–3.50	8	1.8
3.51–4.00	5	1.1
More than 4.00	7	1.6
Total:	445	100.0

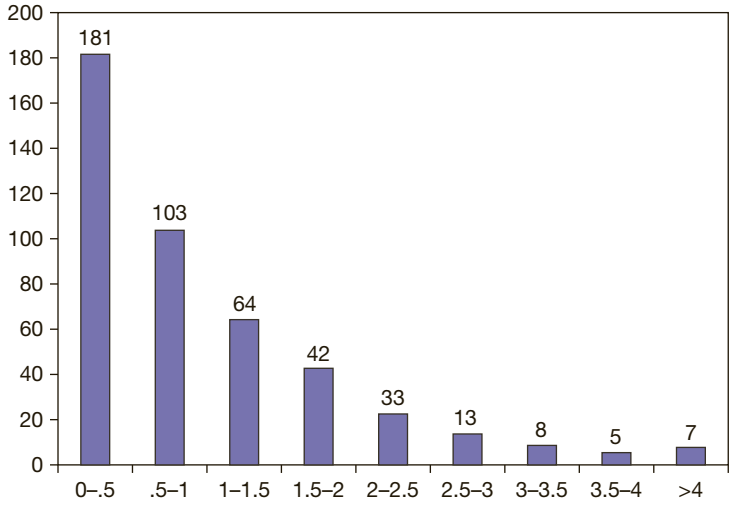


FIGURE 15.3 Length of stay in operating room observed in a sample of 445 patients.

The data in Table 15.2 show that the frequency of length of stay tends to decrease as the length of stay increases. This trend is indicated more sharply in Fig. 15.3 where the frequency is plotted against the length of stay; this kind of graph is called a *histogram*.

Whenever a scientist sees data that shows one variable rapidly diminishing (or increasing) as another variable increases uniformly, he suspects that the two variables are related in an exponential fashion. A statistical analysis of the data of Table 15.2 shows that the distribution of times of surgical procedures closely follows a curve given by what probability theorists would call a *negative exponential distribution* with respect to time. This is a curve that has the form $y = \mu e^{-\mu t}$ where μ is a positive constant representing the reciprocal of the average length of stay in the operating room.

A continuous nonnegative real-valued function whose integral over its entire domain equals 1 is called a *continuous probability density function*. To verify that the positive valued function $f(t) = \mu e^{-\mu t}$ is actually a probability density for $t \geq 0$, we need to show that the improper integral

$$\int_0^\infty f(t)dt = \int_0^\infty \mu e^{-\mu t} dt = 1$$

We evaluate this improper integral using standard calculus techniques:

$$\int_0^\infty \mu e^{-\mu t} dt = \lim_{b\to\infty} \int_0^b \mu e^{-\mu t} dt = \lim_{b\to\infty} -e^{-\mu t}\Big|_{t=0}^{b} = \lim_{b\to\infty} -\left(e^{-\mu b} - e^{-\mu 0}\right)$$

$$= \lim_{b\to\infty} \left(1 - e^{-\mu b}\right) = 1 - 0 = 1$$

In the calculation we have just seen, note that

$$\int_0^b \mu e^{-\mu t} dt = -e^{-\mu t}\Big|_{t=0}^{b} = -\left(e^{-\mu b} - e^{-\mu 0}\right) = \left(1 - e^{-\mu b}\right)$$

We can interpret the value of this integral as the probability that an operation will last at most b hours. For our exponential density function, the probability that an operation lasts at most t hours is given by

$$Pr(t) = 1 - e^{-\mu t}$$

for each $t \geq 0$. More generally, if f is a continuous probability density function on $[0, \infty)$ is associated with some random variable X, then the probability that X takes on a value no larger than b is

$$\Pr(X \leq b) = \int_0^b f(t)dt$$

while the likelihood that X exceeds b is

$$\Pr(X > b) = \int_b^\infty f(t)dt$$

Thus, the probability that an operation lasts between a hours and b hours can be computed as $Pr(b) - Pr(a)$. The *mean* or *average value* of a continuous probability density f on $[0, \infty)$ is $\int_0^\infty tf(t)dt$ if that improper integral converges to a finite value. For more details of these kinds of probability density, the reader may consult Grimmett and Stirzaker [2002] or any other standard probability text.

Schmitz and Kwak assumed that the length of stay in the operating room could be given accurately by a negative exponential distribution with constant μ, obtained by taking the observed average length of stay (1.03 hours) in their sample of 445 surgical cases. This distribution predicts, for example, that 9.0 percent of the operations will last between 1.5 and 2 hours. This compares well with the observed frequency of 9.4 percent in the sample.

B. Random Selection of Patients

We come now to the key step in the simulation process. We are not going to perform any actual operations. Rather, we will select "patients" at random and keep track of how long the patient's operation should last. Our patients will be the 1,000 integers between 000 and 999, inclusive. Since 2% of the patients need operations lasting more than 4 hours (this is determined from the negative-exponential distribution), we must reserve a block of 20 of the integers to represent these cases. Let us say that we reserve the block from 980 to 999. If the number we select at random falls between these limits, then we pretend to perform an operation lasting more than four hours. If the number selected falls outside these limits, then it belongs to a block of numbers representing patients who require an operation for a different length of time.

In Table 15.3, we list the type of surgery, associated length of time, frequency of length of time, and the bloc of numbers reserved for it.

One problem remains. Once we determine that a patient is to be in the operating room for a time, say, between 1.51 and 2.00 hours, how long do we actually keep him there?

Table 15.3 Assignment of random numbers

Type of surgery	Time interval	Relative frequency	Random number block
ENT	0.01–0.50	15.8	000–157
Urology (to RR)	0.01–0.50	08.4	158–241
Urology (no RR)	0.01–0.50	08.5	242–326
Ophthalmology (no RR)	0.01–0.50	05.8	327–384
All other surgery	0.51–1.00	23.6	385–620
"	1.01–1.50	14.6	621–766
"	1.51–2.00	09.0	767–856
"	2.01–2.50	05.5	857–911
"	2.51–3.00	03.4	912–945
"	3.01–3.50	02.1	946–966
"	3.51–4.00	01.3	967–979
"	More than 4.0	02.0	980–999

Table 15.4 Simulated length of operations

Random number	Simulated time in OR (in hours)
000–384	.5
385–620	.75
621–766	1.25
767–856	1.75
857–911	2.25
912–945	2.75
946–966	3.25
967–979	3.75
980–999	4.15

Evidence shows that the average length of an operation requiring less than one-half hour is .47 hours; we will round this number off to .5 for the simulation. For the other categories, we will take as the average length of time the midpoint of the time interval. Thus, a we arrive at the simulated lengths of operating-room time listed in Table 15.4. Schmitz and Kwak actually used a more sophisticated approach, based on the negative exponential distribution, to arrive at the average length of operation for each time interval, but their numbers are not significantly different from ours.

How are the random numbers generated? We wish to ensure that each of the 1,000 numbers 000 to 999 has the same probability of selection. Tables of 1,000,000 or more random integers have been published, and these tables are often used for such simulations. Another method of choosing the numbers would be to build a balanced roulette wheel with a thousand numbered slots on it evenly distributed about the circumference. A random number is then determined by spinning the roulette wheel. Because the roulette wheel is a

familiar device for generating random events, the type of model we are using is often called a *Monte Carlo simulation* in honor of the famous European gambling casino. Many computer languages have subroutines that produce numbers sufficiently uniformly distributed and random to use in simulation experiments; these subroutines are, in fact, simulations of a roulette wheel.

In addition to operating room demands, Schmitz and Kwak were concerned with needs for recovery-room beds. After a surgical procedure in which a general anesthetic is administered, patients are taken to a recovery room in which nurses are constantly present to monitor their vital signs for some period before the patients go back to their hospital rooms. The method by which recovery-room demands are handled in the simulation is explained in the next section, in which we describe in detail the rules of the simulation.

V. The Simulation Model

A. Rules of the Simulation

In carrying out the simulation of the length of stay in the operating room and the recovery room, a set of rules was formulated by Schmitz and Kwak to reflect the medical policies of the hospital. The rules were these:

1. Twenty-seven cases were simulated based on the increased bed complement.
2. The random numbers used to select patients were generated independently for each simulated day.
3. All ENT, urology, and ophthalmology surgical cases have an average length of stay in the operating room of .5 hours.
4. Fifty percent of the urology surgical cases do not go to the recovery room, because they are performed under a local anesthetic. Whether or not a urology case goes to the recovery room is governed by the random number chosen; see Table 15.3.
5. All ENT surgical cases go to the recovery room.
6. None of the ophthalmology cases go to the recovery room. The few ophthalmology cases that actually go to the recovery room in practice are balanced out by the few ENT cases that do not go to the recovery room.
7. Any operation lasting more than .5 hours is considered major surgery and the patient spends 3 hours in the recovery room. Otherwise, if a patient goes into the recovery room, he stays there for 1.5 hours.
8. The starting time for the beginning of the surgical schedule is 7.50—that is, 7:30 a.m.
9. The necessary "make-ready" time from the moment that one surgical case leaves the operating room until it is ready to receive the next case is .25 hours.
10. It takes .08 hours to transport a patient from the operating room to the recovery room.
11. The necessary "make-ready" time from the moment that a patient leaves the recovery room until his bed is ready for the next occupant is .25 hours.

12. The first operating room to be vacated is the first one to be put back into use when the need arises.
13. The first recovery-room bed to be vacated is the first one to be put back into use when the need arises.
14. If there is no previously vacated recovery-room bed when the patient arrives from surgery, a new bed is created.

The only missing piece of information that is necessary to begin the simulation is the number of available operating rooms. Since this is one of the factors to be determined, the simulation may be run for many hypothetical days with different numbers of operating rooms.

B. Results of the Simulation

We illustrate the results of a single simulation with five operating rooms. Table 15.5 contains the necessary information for one simulated day.

We can easily trace through the first steps in constructing Table 15.5.

1. The first random number selected is 889. From Table 15.4, we see that this patient will have an operation lasting 2.25 hours, because 889 belongs to the block of numbers 857−911.
2. Adding the simulated length of the operation to the starting time of the operation (7.50), we find that the operation ends at 9.75. Operating room 1 will then be ready for a new patient at $9.75 + .25 = 10.00$ hours.
3. We add .08 to the ending time of the operation to determine that the patient arrives at the recovery room at 9.83.
4. Since this patient underwent major surgery, he will remain in the recovery room for three hours, leaving it at 12.83.
5. The recovery-room bed he occupied will be ready for another patient at $12.83 + .25 = 13.08$.
6. Since there are five operating rooms in this simulation, the first five patients will start surgery at the same time, 7.50. The second patient on the schedule, represented by random number 396, is the first patient to reach the recovery room, so he is assigned **RR** bed 1. The first patient of the day on the schedule (random number 889) is actually the seventh patient to reach the recovery room. That explains why he is assigned **RR** bed 7.
7. Note also from this simulation that when operating room 1 is ready for its second procedure, it receives the fourteenth patient (random number 648) on the schedule. Surgery for the patients higher on the schedule takes place in the other operating rooms.
8. The third scheduled operation was performed on random number 358. From Table 15.3, we note that this patient is an ophthalmology case and will not go to the recovery room.

9. Random number 214 represents the seventeenth scheduled operation of the day. Table 15.3 indicates that this patient will receive a urology operation requiring 1.5 hours in the recovery room (since 214 belongs to the bloc 158−241).

For this simulated day, using five operating rooms, we discover that the surgical schedule is completed at 14.40 (about 2:24 p.m.), the last departure from the recovery room occurs at 17.73 (about 5:44 p.m.), and that 12 recovery-room beds were needed.

Table 15.5 An example of the simulation

Schedule Number	Random Number	Time Length of Operation	Time Operation Begins	Time Operation Ends	Operating Room Number	Recovery Room Yes No	Time Recovery Begins	Time Recovery Ends	RR Bed No.	Time RR Bed Available
1	889	2.25	7.5	9.75	1	X	9.83	12.83	7	13.08
2	396	.75	7.5	8.25	2	X	8.33	11.33	1	11.58
3	358	.5	7.5	8	3	X	—	—	—	—
4	715	1.25	7.5	8.75	4	X	8.83	11.83	3	12.08
5	502	.75	7.5	8.25	5	X	8.33	11.33	2	11.58
6	68	.5	8.25	8.75	3	X	8.83	10.33	4	10.58
7	604	.75	8.5	9.25	2	X	9.33	12.33	5	12.58
8	270	.5	8.5	9	5	X	—	—	—	—
9	228	.5	9	9.5	4	X	9.58	11.08	6	11.33
10	782	1.75	9	10.75	3	X	10.83	13.83	4	14.08
11	379	.5	9.25	9.75	5	X	—	—	—	—
12	93	.5	9.5	10	2	X	10.08	11.58	8	11.83
13	11	.5	9.75	10.25	4	X	10.33	11.83	9	12.08
14	648	1.25	10	11.25	1	X	11.33	14.33	6	14.58
15	527	.75	10	10.75	5	X	10.83	13.83	10	14.08
16	987	4.15	10.25	14.4	2	X	14.48	17.48	2	17.73
17	214	.5	10.5	11	4	X	11.08	12.58	11	12.83
18	474	.75	11	11.75	3	X	11.83	14.83	1	15.08
19	238	.5	11	11.5	5	X	11.58	13.08	2	13.33
20	45	.5	11.25	11.75	4	X	11.83	13.33	8	13.58
21	408	.75	11.5	12.25	1	X	12.33	15.33	9	15.58
22	116	.5	11.75	12.25	5	X	12.33	13.83	3	14.08
23	209	.5	12	12.5	3	X	12.58	14.08	5	14.33
24	48	.5	12	12.5	4	X	12.58	14.08	12	14.33
25	393	.75	12.5	13.25	1	X	13.33	16.33	11	16.58
26	550	.75	12.5	13.25	5	X	13.33	16.33	7	16.58
27	306	.5	12.75	13.25	3	X	—	—	—	—

C. Conclusions

In their paper, Schmitz and Kwak present the results of four simulated days. On three of the days, 11 recovery-room beds were needed and on the fourth, there was a demand for 12 beds. On all four days, the surgical schedule was completed by 5:30 p.m. The latest departure from the recovery room was about 8:36 p.m. These simulations all assumed that there were five operating rooms available.

The simulation of a daily surgical schedule that we have described can be carried out by hand in less than an hour. On a contemporary computer, the simulation can be completed in mere seconds, so that it is possible to repeat the simulation a great many times at a very modest cost.

It should be apparent from our description that it is a simple matter to vary the number of surgical cases for the day as well as the number of operating rooms. If there are only four operating rooms, then we can see that the surgical schedule of 27 procedures will not be completed until the evening hours, while if we increase to six operating rooms, some will stand empty for a good part of the afternoon. To obtain more precise estimates of time, of course, we need only run through the simulation process with these constraints.

Schmitz and Kwak conducted the simulation using 3, 4, 5, and 6 operating rooms. Based on 27 surgical procedures per day, they discovered that the optimum number of operating rooms was found to be 5 and that there would consistently be a need for at least 12 recovery-room beds. They also concluded that it was not necessary to staff the recovery room beyond 9 or 10 p.m. each day.

D. Validation

The effectiveness of any model is measured by how closely its predictions match those actions of the system which are observed in the real world. One way to test our Monte Carlo simulation of operating room and recovery room usage would be to expand the operating room capacity to five rooms when the bed complement is increased by 144 and then simply observe whether the daily surgical schedule works out as well as the simulation says it should. Unfortunately, this could be a very costly testing procedure, both in terms of construction dollars and patient well-being, especially if the assumptions underlying the simulation are poor ones, or if important factors have been omitted from it.

Fortunately, there is an alternative validation procedure. We can determine how well the model simulates a situation for which observed data *already exists*. In our case, we know how the hospital system functioned in 1970 before the addition of new Medical/Surgical beds. We may then perform our simulation for the daily surgical schedule of 1970. We will be testing the validity of the simulation to predict the future by determining how well it simulates the past or present.

Since there were 6,293 procedures in the year 1970, our simulation would call for 17 (6293/365) procedures per day. Suppose that there were three operating rooms available in 1970, but that all the other rules of the simulation were the same. The results of this simulation, using the random numbers of the first 17 patients of Table 15.5, are shown in Table 15.6.

With a simulated daily surgical schedule of 17 cases and three operating rooms, note that all operations have been completed by 16.90 (4:54 p.m.), that there is a need for seven recovery-room beds, and that the latest time that a patient leaves the recovery room is

Table 15.6 Validating the simulation

Schedule Number	Random Number	Length of Operation	Time Operation Begins	Time Operation Ends	Operating Room	Recovery Room Yes No	Time Recovery Begins	Time Recovery Ends	RR Bed No.	Time RR Bed Free
1	889	2.25	7.50	9.75	1	X	9.83	12.83	4	13.08
2	396	0.75	7.50	8.25	2	X	8.33	11.33	1	11.58
3	358	0.50	7.50	8.00	3	X	—	—	—	—
4	715	1.25	8.25	9.50	3	X	9.58	12.58	3	12.83
5	502	0.75	8.50	9.25	2	X	9.33	12.33	2	12.58
6	068	0.50	9.50	10.00	2	X	10.08	11.58	5	11.83
7	604	0.75	9.75	10.50	3	X	10.58	13.58	6	13.83
8	270	0.50	10.00	10.50	1	X	—	—	—	—
9	228	0.50	10.25	10.75	2	X	10.83	12.33	7	12.58
10	782	1.75	10.75	12.50	1	X	12.58	15.58	2	15.83
11	379	0.50	10.75	11.25	3	X	—	—	—	—
12	093	0.50	11.00	11.50	2	X	11.58	13.08	1	13.33
13	011	0.50	11.50	12.00	3	X	12.08	13.58	5	13.83
14	648	1.25	11.75	13.00	2	X	13.08	16.08	7	16.33
15	527	0.75	12.25	13.00	3	X	13.08	16.08	3	16.33
16	987	4.15	12.75	16.90	1	X	16.98	19.98	4	20.23
17	214	0.50	13.25	13.75	2	X	13.83	15.33	6	15.58

19.98 (about 8 p.m.). If these values are close to those observed for a typical day in 1970, this would reinforce the belief that the method of simulation chosen is a valid one. It is interesting to note here that the average length of the 17 operations on this simulated surgical schedule is 1.04 hours, compared with an observed figure of 1.03 for the sample of 445 cases in 1970.

E. Possible Refinements

There are several ways that we could increase the sophistication of this simulation to imitate better the actual operating- and recovery-room usages. Instead of using 30-minute time intervals for our surgical categories, we could have used 10-minute or even 5-minute intervals. This would give a more precise and accurate distribution of lengths of time in the operating room. The determination of recovery-room usage was based on the simplifying assumption that a patient who arrived in the recovery room would spend either 1.5 or 3 hours there, depending on the length of his operation. In fact, the length of time in the recovery room does not take on only these two values. The nature of the surgery performed, the length of time it took, and the age and general state of health of the patient will all be factors in establishing how many hours he will be kept in the recovery room.

The sample data of the 445 patients could have been used to determine a probability distribution for length of time in the recovery room. This distribution then could have been

incorporated into the simulation in much the same way that the negative-exponential distribution for length of stay in the operating room was.

This method of simulation is thus seen to be extremely flexible, as it allows for various levels of sophistication. As Schmitz and Kwak [1972] point out,

> *In general, this method gives a close approximation to reality under conditions when it is not possible to ascertain by observation the operation of a department. It was found to be extremely accurate when it did become possible to observe the operation of the department. The uses of the method are limited only by the imagination and ingenuity of the user.*

F. How Many Days to Simulate?

In the models we first studied in this book—deterministic, axiomatic, and probabilistic—conclusions about the behavior of a real-world system were *deduced* from a mathematical model by the standard techniques of proving theorems and solving equations. When simulation is used to study an actual system, the conclusions we reach can only be *inferred* from the outcome of sample runs of the simulation.

Since chance plays such a basic role in a Monte Carlo simulation, repeating the simulation a second time—with exactly the same rules—will produce different results, because different random numbers will be generated. We cannot be content with running our simulation for one daily surgical schedule and making predictions on the basis of the outcomes we see. We need to repeat the simulation many times to assess the effect of chance on the differences in outcomes that will be produced. But how many times is a sufficient number? Consider a prediction, for example, that 12 recovery-room beds will be sufficient for five operating rooms. How does the degree of confidence in this prediction grow with the number of simulated days on which no more than 12 beds are demanded? The proper answer to questions like this and for the general evaluation of simulation experiments may require quite sophisticated statistical techniques. Some problems were solved quite a while ago (see Chapter 8 of John Smith, *Computer Simulation Models*), but many thorny difficulties remain. Mathematicians are actively developing a theory of simulation that will enable this powerful technique to be used more widely and knowledgeably.

For the simulation of operating-room and recovery-room usage, we would like to have a firm grasp both of the expected number and the variance of recovery-room beds would be needed for each possible number of operating rooms. As we noted, repetition of the simulation many times may be necessary to achieve such an understanding. It is important that each simulated day be as independent as possible from any other day. We need then to have a new set of random numbers for each simulated run. How do we obtain strings of random numbers?

When simulation and the Monte Carlo method began to gain popularity in the 1950s, a need arose for a large source of numbers that were as random as possible. A single simulation of a complex process might easily use hundreds of thousands of random numbers. The RAND corporation responded with a book, still in print, titled *A Million Random Digits with 100,000 Normal Deviates*. To generate truly random numbers, RAND essentially built an electronic roulette wheel. RAND's tables have been widely used in engineering, econometrics, statistics, public opinion polling, physics, and lotteries. Other attempts to produce random numbers have used emissions from radioactive materials or other natural processes.

Many simulation studies today find it to burdensome to store and retrieve previously generated random number sets. It is faster to have the computer generate a "random" number with a simple algorithm whenever one is needed. The quotes around "random" are deliberate. The computer is a deterministic device; the same input should always produce the same output. The computer is not capable of producing lists of numbers in a truly random fashion. To deal with this limitation, scientists have developed algorithms that are fast and efficient and that yield strings of numbers that pass standard statistical tests for randomness (see Exercise 23 for an example). We call these deterministic algorithms *pseudorandom number generators*.

Pseudorandom number generators were used even in the first generation of simulations run on computers with very limited memory and relatively slow input and output from punched cards. Von Neumann considered hardware-based true random number generators inappropriate either because they kept no record of numbers generated preventing later tests for errors or because if they stored the results, the numbers quickly exhausted computer memory. He advocated rapid simple algorithms, such as the Middle Square method (Exercise 22), cautioning that their output should not be confused with truly random sequences: "Anyone who considers arithmetical methods of producing random digits is, of course, in a state of sin," he quipped.

The Middle Square Method was fairly quickly replaced by more sophisticated algorithms, which use formulas to generate the next number in a pseudorandom sequence by performing complicated arithmetic functions on the current number. Few, if any, simulations today employ natural authentically random processes. As mathematician Robert Coveyou asserted, "The generation of random numbers is too important to be left to chance."

VI. Other Examples of Simulation

The Monte Carlo method was first used to solve problems in nuclear physics where more traditional mathematical techniques failed to give the needed numerical results or required too great a period of time to produce a useful answer. Many applications of the Monte Carlo method involve complicated dynamic behavior entailing a chain of events where at each stage there are different probabilities for the possible outcome of an event. The method, however, can also be effectively applied in situations in which there are, at least on the surface, no probabilities involved.

As a very elementary example, suppose you wish to evaluate a particular definite integral

$$\int_a^b f(x)dx$$

where f is a continuous function on the interval $[a, b]$. From the Fundamental Theorem of Calculus, this is a trivial problem provided you can find an antiderivative of f—that is, a function F such that $F'(x) = f(x)$ for all x in $[a, b]$. Then the value of the definite integral is given as the difference $F(b) - F(a)$.

In many instances, the function F cannot be found so easily. In fact, for most functions f (examples are $f(x) = \dfrac{\sin x}{x}$ and $f(x) = e^{-x^2}$), it is impossible to find F in closed form as a rational combination of the standard functions of calculus.

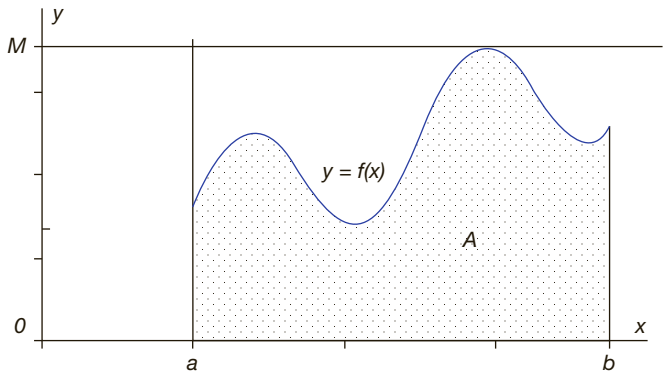


FIGURE 15.4 The shaded region has area $A = \int_a^b f(x)dx$.

There are various approximation techniques in such cases for finding the numerical value of the definite integral. The Monte Carlo method is one such technique.

For convenience, suppose that the function f takes on only nonnegative values and that it is bounded on the interval $[a, b]$ by the positive number M. Then the graph of f over the interval is entirely contained in a rectangle of dimensions $(b - a)$ by M. (See Fig. 15.4.)

The value of the definite integral $\int_a^b f(x)dx$ is the measure of the shaded area A in the rectangle that is below the graph of the curve $y = f(x)$. The relative area,

$$p = \frac{A}{(b-a)M}$$

is then a number between 0 and 1. This number can be interpreted as a probability. If a point is picked completely at random from the points of the rectangular region, then the probability that the point lies in the shaded area under the curve is precisely the number p of Eq. (1).

If there is some other independent way of finding the probability p, then the value of the definite integral can be found simply as

$$\int_a^b f(x)dx = A = p(b-a)M \tag{1}$$

The Monte Carlo method provides a way of obtaining the probability p directly. We need to recall the *relative frequency* interpretation of probability introduced in Chapter 10. Imagine the experiment of choosing a point at random from the rectangular region and noting whether or not it lies in A. If this experiment is repeated a large number N of times, then p is approximately the frequency of selecting points in A—that is,

$$p \approx \frac{\text{Number of times point chosen lies in } A}{N} \tag{2}$$

Our understanding of probabilities is that the approximation in Eq. (2) improves as N increases. If we conduct the experiment 10,000 times, then we should get an accurate estimate of p and hence an accurate estimate of $\int_a^b f(x)dx$.

A point can be chosen at random from the rectangular region by first choosing its x-coordinate at random and then choosing its y-coordinate at random. The x-coordinate must lie between the numbers a and b and the y-coordinate between 0 and M. If we have a random number generator that produces numbers between 000 and 999 with an equiprobable distribution, then we may start by generating two random numbers, r and s. We then determine a point in the desired region with coordinates (x_0, y_0) where

$$x_0 = a + \frac{r(b-a)}{999}$$

and

$$y_0 = \frac{sM}{999}$$

To decide whether the randomly chosen point (x_0, y_0) belongs to A, we need only compute $f(x_0)$ and determine whether the inequality

$$y_0 < f(x_0) \tag{3}$$

is valid. If it is, then the point belongs to A; otherwise, it does not. To obtain our approximation for p, we generate N points in the rectangular region in the manner just described and keep track of what proportion of times the coordinates of the point satisfy the inequality (3).

Example

Let's start with an example of an integral whose value we do know and see how well the Monte Carlo method works. Examine $f(x) = x^2$ on $[0, 1]$ where

$$\int_0^1 x^2 dx = \frac{x^3}{3}\Big|_{x=0}^{x=1} = \frac{1}{3} - \frac{0}{3} = \frac{1}{3} = 0.3333\ldots$$

Table 15.7 shows the results of 10 simulations, each using 10,000 randomly selected points.

Note that each estimate for the value of the integral, based on randomly choosing 10,000 points, is close to the true value of 1/3. Moreover, the average value for the 10 simulations is .33321. My computer carried out all 10 simulations in a fraction of a second.

For a second example, we'll use an integral we can't evaluate via the Fundamental Theorem of Calculus.

Table 15.7

Simulation Run	Monte Carlo Estimate of $\int_0^1 x^2 dx$
1	0.3314
2	0.3343
3	0.330
4	0.3357
5	0.333
6	0.3363
7	0.3391
8	0.3332
9	0.3239
10	0.3352

Example

Use the Monte Carlo technique to estimate $\int_0^1 \frac{\sin x}{x} dx$. Table 15.8 shows the results of 10 simulations, each one involving choosing 1,000,000 points at random. On a small laptop computer, the calculations took less than 90 seconds to complete.

The average of the 10 estimates is 0.9458502. The definite integral $\int_0^1 \frac{\sin x}{x} dx$ can be estimated by other techniques such as the use of Riemann sums or the Trapezoidal Rule. These estimates give a value of 0.9460830704.

Table 15.8

Simulation Run	Monte Carlo Estimate of $\int_0^1 \frac{\sin x}{x} dx$
1	0.9458610000
2	0.9456010000
3	0.9456070000
4	0.9460810000
5	0.9460680000
6	0.9459830000
7	0.9458220000
8	0.9458780000
9	0.9458270000
10	0.9457740000

The real value of the Monte Carlo method comes in situations where other techniques to obtain approximate numerical answers either do not exist or are much less efficient. Not all simulations use the Monte Carlo device. "Deterministic" simulation has been used very successfully in studying the path of a spaceship, for example. If a rocket is sent on a lunar landing expedition, the system consists of *four* bodies exerting gravitational attractions on each other: the sun, earth, the moon, and the spaceship. Newton's laws give the deterministic differential equation controlling the path of flight of the ship. We can simulate the solution curve of the equation by using a sophisticated version of the Euler method discussed in Chapter 2.

VII. Historical and Biographical Notes

The Monte Carlo method has a humble origin. In 1946, the Polish-American mathematician Stanislaw Ulam was recovering in Los Alamos from a severe case of encephalitis, an acute inflammation of the brain. Advised by his doctors not to think too strenuously, he wiled away some time playing a solitaire card game and began to wonder how often the arrangement of cards might result in a win:

> *I noticed that it may be much more practical to get an idea of the probability of the successful outcome . . . by laying down the cards . . . and merely noticing what proportion comes out successfully, rather than try to compute all the combinatorial possibilities which are an exponentially increasing number so great that, except in very elementary cases, there is no way to estimate it. This is intellectually surprising, and if not exactly humiliating, it gives one a feeling of modesty about the limits of rational or traditional thinking. In a sufficiently complicated problem, actual sampling is better than an examination of all the chains of probability.*

Ulam quickly realized that such an approach could be used to study all processes involving branching of events, including one of special interest to the atomic scientists at the time: the production and further multiplication of neutrons in some kind of material containing uranium or other fissile elements. A neutron might scatter at one angle, change its velocity, be absorbed, or produce more neutrons by fission, Ulam noted. The elementary probabilities for each of these possibilities were individually known, but

> *The problem is to know what succession and branching of perhaps hundreds of thousands or millions will do. One can write . . . equations for the "expected values," but to solve them or even get an approximate idea of the properties of the solution, is an entirely different matter.*

Ulam's idea, the core of the simulation process, is to try out thousands of scenarios, at each stage selecting a random number with appropriate probability to decide which branch to follow. "After examining the possible histories of only a few thousand," Ulam declared, "one will have a good sample and an approximate answer to the problem. All one needed was to have the means of producing such sample histories, It so happened that computing machines were coming into existence, and here was something suitable for machine calculation."

In addition to his work on the Manhattan Project to develop the first atomic bomb in World War II, Ulam (April 13, 1909−May 13, 1984) was an outstanding pure and applied

mathematician, originating the Teller-Ulam design of thermonuclear weapons. The term "Monte Carlo" was apparently coined by the Greek-American physicist Nicholas Metropolis (June 11, 1915−October 17, 1999), who also worked at the Los Alamos laboratory and originated several algorithms for generating random numbers to use in these simulations.

Although simulation was initially employed in the physical sciences, by the mid-1950s social scientists were beginning to apply it also to a variety of problems. Simulation has been used as a tool for research, teaching, decision-making, and historical reconstruction. Social scientists have used simulation to study specific topics such as the spread of urban ghettoes, the outbreak of World War I, the behavior of the stock market, the introduction of a new product in a competitive market, and neurotic processes in psychopathology. Disciplines as diverse as geography, political science, cognitive and social psychology, medicine, international relations, anthropology, education, sociology, and business administration have all been affected to some extent by the results of simulation studies.

A characteristic of discrete event simulation, as you have seen in our hospital planning model is viewing a system as a sequence of events that occur at particular instants of time and signal a change of state in the system. The simulation jumps in time from one event to the next one in the sequence; the system does not change in the interim between events.

An alternative approach, called *continuous simulation*, breaks up the time interval of interest into small time periods or "slices" and updates the system at the start of the next period based on the activities that occurred since the beginning of the previous period. Since no change may occur during a specified time slice, continuous simulation may run more slowly than a discrete-event approach, which does not have to simulate every period.

Contemporary simulations can easily transform the numerical data generated by the underlying equations or rules that govern the simulation into pictures, graphs, or animations. In watching these visual displays of the output of a simulation, the viewer can obtain an impressive qualitative sense of the simulation's behavior. Sometimes, one may be seduced into believing that what is being displayed is the actual behavior of the real-world system. As Peter Bak warns in his book *How Nature Works*, New York: Copernicus, 1996:

> *There is no such thing as doing calculations on the real thing. One cannot put a frog into the computer and simulate it in order to study biology. Whether we are calculating the orbit of Mercury circling the sun, the quantum mechanics of some molecule, the weather, or whatever, the computer is only making calculations on some mathematical abstraction originating in the head of the scientist. We make pictures of the world. Some pictures are more realistic than others. Sometimes we feel that our modeling of the world is so good that we are seduced into believing that our computer contains a copy of the real world, so that real experiments or observations are unnecessary. I have fallen into that trap when sitting too long in front of the computer screen.*

For a more complete introduction to this active and growing subject, you may wish first to examine the book by Guetzkow and others [1972], which contains essays on the advantages and limitations of simulation as well as a number of particular case studies from the relatively early days of simulation studies. More recent work can be found in the works of Gilbert [1999]. One of the most authoritative texts on the theory and practice of simulation is Averill Law's book [1999].

Homer H. Schmitz currently serves as a professor of health management and policy at the College of Public Health and Social Justice of Saint Louis University, where he has been a faculty member since 2002. At the time he and Professor Kwak developed this simulation model, he was the vice president and director of management services at Deaconness Hospital. Professor Schmitz has extensive executive experience in managing the operations, information systems, planning, and finances of various sectors of the health care market including a 450-member multispecialty physician practice, a managed care organization with over 250,000 enrollees, an emergency medical service organization with over 100 vehicles, and a 500-bed acute care teaching hospital.

An internationally recognized author and lecturer in health care management, Professor Schmitz has authored or coauthored five books and more than 75 articles. He has also had a number of consulting assignments throughout the United States, Syria, the United Arab Emirates, Qatar, and South Africa.

Photo courtesy of Homer Schmitz

Homer H. Schmitz

Photo courtesy of N. K. Kwak

NoKyoon Kwak

NoKyoon Kwak is an emeritus professor of decision sciences at the John Cook School of Business of Saint Louis University.

Author or coauthor of more than 125 journal articles, Professor Kwak has written a number of books: *Management Science: Theory and Applications*, *Introduction to Mathematical Programming*, *Operations Research: Applications in Health Care Planning*, *Managerial Applications of Operations Research*, *Quantitative Models for Business Decisions*, *Quantitative Decision Theory for Management*, and *Mathematical Programming with Business Applications*. Several of these have been translated into Chinese and Korean.

He also served as a Fulbright distinguished visiting professor at Dongguk University in Seoul, Korea, and at Beijing Agricultural Engineering University and Shengyang Agricultural University in the People's Republic of China. His most recent research has included studies on using data envelopment analysis for performance comparison of Missouri public schools and using multicriteria decision-making models in resource planning for health care systems and for strategic outsourcing and supply chain management.

EXERCISES

1. The projected daily surgical schedule of 27 procedures was based on the assumption that operations would be performed every day. Find the length of the schedule if no operations are done on Sunday.

2. Of the 6,293 surgical procedures performed in 1970, approximately how many were ophthalmology cases?

3. Find the annual total number of ophthalmology surgical procedures projected for the expanded bed complement.

4. Find the percentage of the projected 9,669 procedures that will be ophthalmology cases.

5. Compare the percentage obtained in Exercise 4 with the 4.5% experienced in 1970. Are the percentages the same? Should they be the same?

6. Suppose that 6% of the surgical procedures in 1970 were gynecology cases and that the average stay in the hospital for such a case was 4.7 days. Show that this would yield a projected increase of 279 gynecological procedures for the expanded bed complement.

7. If a projected increase of 312 beds is estimated for a surgical category that represented 7.1% of all procedures in 1970, what was the average length of stay per patient in this category?

8. Schmitz and Kwak report that the projected 9,669 annual procedures would represent an increase of 53.6% over the 1970 totals. Does this give you enough information to determine the number of MIS beds in the hospital in 1970? To determine the average length of stay of a surgical patient? What other data would you need in order to answer these questions?

9. For a random variable X with continuous probability density function f defined on $[0, \infty)$, show that

(a) The probability that X takes on a value larger than b is $\int_b^\infty f(t)dt$.

Hint: Use the fact that $\int_0^\infty f(t)dt = \int_0^b f(t)dt + \int_b^\infty f(t)dt$.

(b) $\Pr(a \le X \le b) = \int_a^b f(t)dt$.

10. For the negative exponential density $f(t) = \mu e^{-\mu t}$, show that the probability that an operation will last more than T hours is $e^{-\mu T}$.

11. **(a)** Use integration by parts to show that $\int \mu t e^{-\mu t} dt = \frac{-(1+\mu t)e^{-\mu t}}{\mu} + C$.

(b) Suppose that $\mu > 0$. Use the result of part (a) to show that $\int_0^\infty \mu t e^{-\mu t} dt = \frac{1}{\mu}$. (l'Hôpital's Rule may be helpful.)

Exercises 12–19 refer to the simulation of Table 15.5.

12. In which operating room was the largest number of procedures performed? The smallest number? What was the average number of procedures per operating room?

13. Which operating room was used for the longest period of time? What was the average length of time of usage per operating room?

14. Using the random numbers of Table 15.5, trace through the effects on the number of required recovery-room beds of

(a) shortening the "make-ready" time in the recovery room to .20 hours;

(b) lengthening the recovery-room time to 4 hours for major surgery and 2 hours for minor surgery.

15. Using the random numbers from Table 15.5, work through the simulation with the original rules, but with only three operating rooms. Determine the following:

(a) Time of completion of surgical schedule

(b) Number of recovery-room beds required

(c) Latest time a patient leaves the recovery room

16. Repeat Exercise 12 with

(a) Six operating rooms

(b) Four operating rooms

17. Trace through a simulated surgical schedule using the original rules and the random numbers of Table 15.5, except that whenever a random number R occurs, choose the patient represented by the random number $999 - R$.

18. Generate your own random number sequence and go through a simulated day.

19. What refinements in the model would you suggest that would make the simulation more realistic? For some hints, see Fetter and Thompson [1965].

20. Use the Monte Carlo method to estimate $4\int_0^1 \frac{1}{1+t^2}\,dt$. Explain why you expect to get a number close to π.

21. Use the Monte Carlo method to estimate $\int_{-10000}^{10000} e^{-t^2}\,dt$.

22. One of the earliest suggested schemes for generating a sequence of numbers that may appear to be random was the Middle Squares method. Begin with an arbitrary four-digit integer, square it and extract the middle four digits for the next number (if the square has fewer than eight digits, pad some zeroes onto the left end). The third number in the sequence would be the middle four digits of the square of the second number, and so forth.

(a) Show that if the initial four-digit number is 6543, then the next numbers generated are 8108, 7396, 7008, 1120.

(b) The Middle Squares method produces a sequence of numbers between 0 and 9999. How would modify these numbers to get *probabilities*—i.e., numbers between 0 and 1?

(c) What does the Middle Squares method produce if you begin with 2500?

(d) What weaknesses in the Middle Square method are revealed by the result in (c)?

(e) Show that the middle four digits can be extracted by first dividing the square by 100 and throwing away the decimal part and then dividing the resulting number by 10,000 and keeping the remainder.

23. The Linear Congruential Generator is a widely used method for producing strings of numbers that appear random. It uses the relationship $x_{n+1} = (ax_n + c) \bmod m$ to produce the next number x_{n+1} in the sequence from the previous number x_n and fixed integers a, c, and m. After computing $a\,x_n + c$, divide it by m and let x_{n+1} be the integer remainder.

(a) If $a = 231$, $c = 13$, and $m = 2^{10}$, show that an initial choice of $x_1 = 1221$ produces a sequence beginning 404, 701, 152, 309, 736, . . .

(b) Java uses $a = 25214903917$, $c = 11$, and $m = 2^{48}$. What are the first few terms generated by this rule if the initial choice x_1 is again 1221?

SUGGESTED PROJECTS

1. Investigate the negative exponential distribution. Show, in particular, that the simulated length of stay in the operating room can be determined by drawing a number from a list of random exponential numbers of mean 1 and then multiplying it by an appropriate interarrival mean. See Schmitz and Kwak's paper and Grimmett and Stirzaker's book (References).

2. (For those with some background in statistics.) Suppose that the simulation we have described is repeated N times and that it is observed that the largest number of recovery-room beds ever needed is 12 and that the latest time an operation was completed was 6 p.m. How large should N be so that we can assert that the probability of needing more than 12 recovery-room beds or that an operation would continue past 6 p.m. is less than .051 (see Chapter 8 of John Smith, *Computer Simulation Models*)?

3. Write a computer program to carry out the Monte Carlo simulation of operating-room and recovery-room usage described in this chapter. Carry out the simulation for a 30-day period and analyze the results.

4. Show that the data usually collected in a major league baseball game to determine players' batting, pitching, and fielding averages gives sufficient information to construct a Monte Carlo simulation that provides estimates on a team's run production as a function of the particular batting order chosen. Discuss the relative difficulty of modeling some particular aspect of football, basketball, or hockey by simulation. What information would be required? Is it readily available?

5. Suppose you are the manager of a supermarket. You must decide the maximum number of items to allow a customer to bring through the express lane. You wish

to select the number that will minimize the average length of time all customers must wait in line before being checked out. What data would you need? How would you construct the simulation?

6. Write a computer program to evaluate definite integrals using the Monte Carlo technique. Test the program on functions whose integrals can be computed exactly by Fundamental Theorem of Calculus. How many points need to be chosen to obtain a good approximation? How does the Monte Carlo method compare in efficiency to other techniques, such as Simpson's Rule?

7. Investigate methods of generating pseudorandom numbers in use today. What are the strengths and weaknesses of the various techniques. What tests should a sequence of numbers need to pass in order to be considered "random enough" to use in simulations? Useful places to start include Barker [2012], Luby [1996], and Knuth [1997].

You can find a listing of references and suggestions for additional reading on the book's website, www.wiley.com/college/olinick

CHAPTER 16

Game Theory

Games combining chance and skill give the best representation of human life. . . . It would be desirable to have a complete study made of games, treated mathematically.

—Gottfried Wilhelm von Leibniz

I. Two Difficult Decisions

We will begin with two classic tales of characters facing difficult decisions involving love, life, and death. One is the biblical patriarch Abraham and the other Florio Tosca, the title character of Giacomo Puccini's famous opera *Tosca*. We will examine how game theory, a mathematical field created in the 20th century, provides insight into their ultimate behavior.

A. Abraham

The Binding of Isaac (in Hebrew, עֲקֵידַת יִצְחָק, Akedat Yitzhak) is one of the most dramatic and troubling stories of the Old Testament. God tests Abraham by demanding a human sacrifice.

Chapter 22 of the book of Genesis begins:

> *Some time later God tested Abraham. He said to him, "Abraham!*
>
> *"Here I am," he replied.*
>
> *Then God said, "Take your son, your only son, whom you love—Isaac—and go to the region of Moriah. Sacrifice him there as a burnt offering on a mountain I will show you."*

God has directed Abraham to commit an almost unthinkable act: willingly kill his son. Should he comply with God's directive, or should he refuse? What would be the consequences of Abraham's decision? How would God respond to the action Abraham does or does not carry out?

It is abhorrent to us to think of any parent deliberately killing, or even injuring, his or her child. For Abraham, the slaying of his only son bears additional anguish as Isaac was seen to be the next link in a long chain of Abraham's descendants who were to be recipients of God's benevolence. Earlier chapters of Genesis tell of the unique, personal relationship

that has developed between God and Abraham. In several passages, God promises Abraham and his descendants great rewards:

> *Raise your eyes and look from where you are, to the north and south, to the east and west. For I give all the land that you see to you and your offspring forever. I will make your offspring as the dust of the earth, so that if one can count the dust of the earth, then your offspring too can be counted. Up, walk about the land, through its length and its breadth, for I give it to you." [Genesis 13:15]*

God reassures him when Abraham, now over 75 years old, expresses fear that he will die childless and his steward will inherit his estate:

> *None but your very own issue shall he your heir. . . . Look toward the heaven and count the stars, if you are able to count them. . . . So shall your offspring be. [Genesis 15:4–5]*

On a third occasion, God reiterates his promise to the now 99-year-old Abraham:

> *I will place My covenant between Me and between you, and I will multiply you very greatly. . . . As for Me, behold My covenant is with you, and you shall become the father of a multitude of nations. . . . And I will make you exceedingly fruitful, and I will make you into nations, and kings will emerge from you. . . . And I will establish My covenant between Me and between you and between your seed after you throughout their generations as an everlasting covenant, to be to you for a God and to your seed after you.*

Abraham and his beloved wife Sarah have lived together for many decades, but they have been unable to conceive a child, for she is barren. When Abraham is 99 years old and Sarah not much younger, it seems that it is impossible that they will ever become parents. As we read in verse 11 of Chapter 18: "Now Abraham and Sarah were old, and well stricken in age; it had ceased to be with Sarah after the manner of women." But there is another message from God that Sarah will bear a son within the next year. When the promise comes true, there is rejoicing in Abraham's camp, and when Isaac is weaned Abraham prepares a great feast.

It comes as a shock to the reader of the Bible that a scant few pages after this happy event, we encounter the demand from God that Abraham sacrifice the long-awaited child.

We can imagine that Abraham may have spent a sleepless night, but by morning he seems to have made his decision:

> *Early the next morning Abraham got up and loaded his donkey. He took with him two of his servants and his son Isaac. When he had cut enough wood for the burnt offering, he set out for the place God had told him about. On the third day Abraham looked up and saw the place in the distance. He said to his servants, "Stay here with the donkey while I and the boy go over there. We will worship and then we will come back to you."*
>
> *Abraham took the wood for the burnt offering and placed it on his son Isaac, and he himself carried the fire and the knife. As the two of them went on together, Isaac spoke up and said to his father Abraham, "Father?"*
>
> *"Yes, my son?" Abraham replied.*
>
> *"The fire and wood are here," Isaac said, "but where is the lamb for the burnt offering?"*

Abraham answered, "God himself will provide the lamb for the burnt offering, my son." And the two of them went on together.

When they reached the place God had told him about, Abraham built an altar there and arranged the wood on it. He bound his son Isaac and laid him on the altar, on top of the wood.

Then he reached out his hand and took the knife to slay his son.

Will Abraham indeed go forward with this human sacrifice? Will God intervene to stop him, or will He sit back and allow it to happen? Will Abraham decide at the last moment that he cannot kill Isaac and that he must defy God whatever punishment he might suffer?

We will pause at this dramatic moment in the story and introduce another famous account of a difficult decision.

B. Tosca

Puccini's opera *Tosca* has been described as a "tragic love story and a nail-biting thriller, from the famous dark opening chords to its unforgettable conclusion." We are in Rome in June 1800. The singer Florio Tosca is fiercely in love with the painter Mario Cavaradossi. He is in trouble with the law as he helped to hide an escaped political prisoner and fellow revolutionary Angelotti. The local police chief Scarpia lusts after Tosca, who is repulsed by his advances. Scarpia suspects that Cavaradossi may be assisting Angelotti and plays on Tosca's jealously in the hopes that she will lead him to Angelotti.

Scarpia's agents fail to find Angelotti, but they do arrest Cavaradossi. Cavaradossi defies Scarpia and denies knowing anything about Angelotti, so Scarpia orders his interrogation—using any means necessary. Despite being tortured, Cavaradossi taunts Scarpia, who orders his immediate execution. At first Scarpia turns a deaf ear to Tosca's pleas for mercy, but then reveals that the price for Cavaradossi's life is Tosca herself. In despair, she sees no way out, despite her revulsion, which only makes her more desirable in Scarpia's eyes. She indicates agreement to have sex with Scarpia in return for saving Cavaradossi's life. Scarpia apparently orders a fake execution and writes out safe conduct passes for Tosca and Cavaradossi. Tosca, nervously looking around the room, perhaps thinking of a way to escape, sees a sharply pointed knife, which she hides behind her. Scarpia seals the passes and then turns to embrace Tosca and satisfy his lust, exclaiming "Tosca, now you are mine at last!"

The moment of truth is hand for Tosca. Should she go forward with her end of the agreement? **Will She Love the Man She Hates to Save the Man She Loves?** How should Scarpia behave? Should he let Cavaradossi go free, or should he double-cross Tosca and make sure his rival dies after he satisfies his carnal desires? Should Tosca double-cross Scarpia, stab him to death with the knife, and flee with Cavaradossi after the fake execution? Will she fight, or will she succumb? The music is building as we sweep toward the end of the opera's second act.

II. Game Theory Basics

A. What Is Game Theory?

Can mathematics help us find a way to advise Abraham and God, Tosca and Scarpia? Are there rational choices for each?

In the dilemmas facing Abraham and God, Tosca and Scarpia, we have a common thread. Several decision makers must make choices that will determine the outcome. No single participant controls the scenario on his or her own.

Much of classical mathematics concerns itself with optimal decision making by a single individual. What shape should a farmer choose to create a pasture with a fixed amount of fencing in order to maximize its area? What is the least expensive meal you can buy at your local fast-food restaurant that meets or exceeds some minimal nutritional requirement? In such situations, there is a solitary decision maker.

Most real-world situations are more complicated. There is an interdependent decision process whose outcome depends on the choices of *all* the actors. In deciding what you should do, you must take into account somehow what everyone else will choose. What should a rational person do? In the early 1940s, John von Neumann and Oskar Morgenstern, a mathematician and economist, respectively, set out to create a new discipline to analyze such situations: game theory. [See Chapter 8 for biographical notes on von Neumann and Morgenstern.]

Up to this point in our text, we have shown modelers as *consumers* of existing mathematics. Richardson, for example, employed systems of differential equations—a mathematical tool invented to study the physical universe—to model the behavior of nations and their weapons. Von Neumann and Morgenstern realized that the mathematics adequate to describe inanimate nature was inadequate to their goals. They had to become *producers* of mathematics.

In an essay written a quarter-century after the publication of their work *Theory of Games and Economic Behavior*, Morgenstern emphasized the inadequacy of the mathematics developed to study the physical world to be useful in studying the social world:

> *Game Theory is a new discipline that has aroused much interest because of its novel mathematical properties* ***and*** *its many applications to social, economic, and political problems. The theory is in a state of active development. It has begun to affect the social sciences over a broad spectrum. The reason that applications are becoming more numerous and are dealing with highly significant problems encountered by social scientists is due to the fact that the mathematical structure of the theory differs profoundly from previous attempts to provide mathematical foundations of social phenomena. These earlier efforts were oriented on the physical sciences and inspired by the tremendous success these have had over the centuries. Yet social phenomena are different: men are acting sometimes against each other, sometimes cooperatively with each other: They have different degrees of information about each other, their aspirations lead them to conflict or cooperation. Inanimate nature shows none of these traits. Atoms, molecules, stars may coagulate, collide, and explode but they do not fight each other: nor do they collaborate. Consequently, it was dubious that the methods and concepts developed for the physical sciences would succeed in being applied to social problems.*

Game theory is the development and analysis of mathematical models of cooperation and conflict among intelligent rational decision makers. Commonly played board games such as chess, checkers, go, Scrabble, and Monopoly, or card games such as bridge, poker, and gin rummy, also involve outcomes that are the result of two or more players' making independent or perhaps coordinated choices. These games contain many of the major concepts in game theory. It is not surprising that the theory borrows many of the terms of ordinary games: players, rules, moves, strategies, payoffs, and so forth. Some of the early work in the theory was motivated by attempts to find optimal strategies in ordinary games.

Although the focus of von Neumann and Morgenstern's book was developing a theory that could be applied to economics, they had a larger vision; at one point, for example, they were thinking of titling their book *General Theory of Rational Behavior*.

B. Classifying Games

Game Theory deals with a vast range of decision-making situations. To make progress on the theory, it is helpful to consider the fact that there are many ways in which games are classified. We will discuss several important classifications.

1. Number of Players

One of the most important classification schemes involves the number of players. There are fields within game theory that deal separately with games involving only one player, exactly two players, three or more players, or extremely large numbers of players.

One-person games are often called *games against nature*; mathematicians use the term *decision theory* to refer to the discipline that considers such situations. The earliest work in game theory, dating from the 1920s, dealt with two-person games. The examples in this chapter focus principally on two-person games. Games with at least three players, called *n-person games*, are especially interesting because of the possibility that subsets of the players may form coalitions who may coordinate their decisions. *Nonatomic games* are situations in which there are an enormous number of players, no single one of whom has very much power or influence—think of a large economy with millions of consumers.

We present here one very elementary example of a one-person game. In keeping with a biblical theme, we will look at David's decision to fight Goliath. Chapter 17 of the book of I Samuel recounts the famous story. The armies of the Israelites and the Philistines face each other on the eve of battle, each occupying a hill with the Valley of Elah between them. From out of the Philistine camp strides the giant Goliath, a warrior nearly 10 feet tall. Each morning for 40 days he taunts the Israelites:

> *"Choose a man and have him come down to me. If he is able to fight and kill me, we will become your subjects; but if I overcome him and kill him, you will become our subjects and serve us. This day I defy the armies of Israel! Give me a man and let us fight each other."*

Saul, the Israelite king, and all his troops are dismayed and terrified; they flee from Goliath in great fear. No one is willing challenge Goliath, despite the promise of rewards if he is triumphant against the giant: "The king will give great wealth to the man who kills him. He will also give him his daughter in marriage and will exempt his family from taxes in Israel."

The young man David, who come to the camp to bring supplies to his oldest brothers serving in the army, hears rumors of these rewards and asks several men independently what good things may be in store for the person who slays Goliath. He is actively thinking about accepting the giant's challenge and weighing the costs and benefits of such a decision.

The best possible outcome for David is that he fights and kills Goliath, marries the king's daughter, becomes a wealthy man whose family is exempt from taxes, and is lauded as the nation's savior and hero. Let W (for win) be the utility David receives from this outcome. See Chapter 8 for more about utility.

At the other end of the spectrum is David's worst outcome: he loses the fight and is slain by Goliath. He may receive some posthumous praise for challenging the giant, but also

anger from a people who are now enslaved as a result of his actions. Let L be the utility of that outcome.

There is an intermediate outcome if David chooses not to fight. He continues to live, perhaps as a poor shepherd with no local fame or fortune. Suppose M is the associated utility so that $L < M < W$.

What is not certain is the outcome of the fight between David and Goliath should the young man accept the giant's challenge. Let p be the probability that David triumphs so that $1 - p$ is the likelihood that Goliath is the victor. One outcome of the fight has value W, which will occur with probability p. The other outcome, with probability $1 - p$, is L. Thus, the expected value to David of the battle would be $pW + (1 - p)L$. Since the option of not fighting gives David a utility of M with probability 1, its expected value is $1M = M$.

Decision theorists might advise David to choose the option with the greater expected value. Thus, David should fight Goliath if and only if

$$pW + (1 - p)L > M$$

which occurs exactly when

$$p > \frac{M - L}{W - L}$$

We can infer from the fact that David did go out and fight Goliath that he estimated his probability p was that large.

2. Zero-Sum versus Nonzero-Sum Games

Our theory concerns itself with games that eventually end. Depending on the choices of players at each turn of the game, and the chance elements that may play a role intermittently, various different outcomes may be possible. Generally there is a dispersal of rewards (money, power, prestige) to each of the players that is dependent on the specific outcome. Von Neumann and Morgenstern posited that there would a *payoff* to each player—a real number measuring the utility each player would receive. They originally thought to confine payoffs to monetary amounts, but quickly realized that a general concept was needed. Thus, they had to begin their theory with creating the axioms of *utility theory* (see Chapter 8).

If there are n players in a game, then a payoff is a n-dimensional vector of real numbers whose ith component is the utility awarded to player i.

If the sum of the entries in every payoff vector of a particular game is 0, then the game is called a *zero-sum* game. An equivalent characterization of a zero-sum game is that the sum of components of every payoff vector is the same constant. A game is a *nonzero-sum* one if there is at least one payoff vector whose components do not add up to 0; equivalently, there are at least two payoff vectors whose sums are different. In a zero-sum game, a player can attain a larger payoff only at the expense of at least one other player suffering a smaller playoff. Compare this idea to Pareto-optimal allocations (see Chapter 9).

Many classic two-person games (chess, checkers, backgammon, and the like) are zero-sum games. These are games of pure conflict. There is no room for negotiation or compromise; they are situations of strict competition.

The *Battle of the Sexes* game is an example of a two-person nonzero-sum game. As described by Anatol Rapoport [1966], a husband and wife are negotiating how they will

spend the evening. The man suggests the opera, the woman suggests a prize fight. Each would rather do something together than not, but they are willing to go their separate ways. Suppose the man gets one utility unit if they both go to the opera, while the woman gets nothing. If they jointly attend the prize fight, the woman gets one, the man nothing. If he goes to the opera and she attends the prize fight, then both receive a negative amount.

Here is a three-person nonzero-sum game. A wealthy man with three daughters dies. His will stipulates that if at least two of them agree on how his estate should be divided, then it will be split according to that agreement. Otherwise, none of them will receive anything, so (0, 0, 0) is a possible payoff vector. Another possible payoff is that each receives one-third of the estate. Suppose that the youngest daughter indeed proposes such an even split: (1/3, 1/3, 1/3). The eldest daughter quickly approaches the middle sister and suggests they agree on splitting the entire estate evenly between the two of them, giving the youngest one nothing: (.5, .5, 0). The youngest quickly comes up with a counteroffer to the middle daughter which she hopes will tempt her: "Let's you and I split it in 60-40; you can have 60%." The proposed payoff vector here is (0, .6, .4). Can you think of a good counteroffer by the oldest sister? Here the eldest and youngest sisters are each trying to build a coalition with the third sister; if two can agree, each of the pair may get more than the one-third share.

3. The Role of Chance

Game theorists also distinguish among decision-making situations by the extent to which chance plays a role in the outcome. At one end of the spectrum are games of pure skill such as chess, checkers, or go, in which chance elements are completely absent. At the other end are games of pure luck, typified by casino offerings like slot machines, roulette, or craps.

Many games fall somewhere in between these extremes. Card games, such as poker, bridge, gin rummy, or blackjack, typify these intermediate situations. Cards are initially shuffled, randomizing their order, and then dealt out to the players so each play of the game begins with a different set of initial conditions over which the players have no control. Once the game begins, however, there are opportunities for more skillful players to do better than less skillful opponents. They may be able to use information about previously revealed cards to adjust their bets (as in blackjack) or compute the probabilities that a specific player holds a particular grouping of cards (as in bridge). Expected value considerations (see Chapter 10) play an important role in analyzing games where chance plays a significant role.

4. Information

Other aspects of a game that determine their character are *information* and *communication*.

Many games involve a sequence of moves made by the players in some rotation. In chess, for example, two players alternate moving a piece from one square on the board to another. In Monopoly, players in turn roll a pair of dice, move a token the indicated number of squares, and then take some action: collect a reward, pay a penalty, buy or upgrade a property, and so on. Although chance plays no role in chess and plays an important role in Monopoly, each player knows exactly what all the other players have done at every move. These are games of *complete information*.

Most card games are decision-making situations with *incomplete information*. In gin rummy, for examples, the two players each receive 10 cards dealt face down from a well-shuffled deck of 52 cards. Thus, neither player knows the initial holding of the other. The

twenty-first card is turned face up to start the discard pile and the remainder of the deck is placed face-down beside it to form the **stock**. Players alternate turns. If it's your turn, then your move has two parts. First, you take the top card from the stock pile or the top card on the discard pile. Both players can see the discard pile, but only you will see the top card from the stock pile if that is the pile you choose. You add whichever card you picked to your hand. For the second part of your move, you remove one card from your hand and place it face up on the discard pile. In gin rummy then, each player has *partial information* about that status of his opponent's hand, but not complete knowledge.

At the other end of the spectrum from chess in terms of information, is the game of Rock-Paper-Scissors. In this game two players simultaneously form one of three shapes with an outstretched hand. The rock beats the scissors, the scissors beat the paper, and the paper beats the rock; if both players throw the same shape, the game is tied. Since the players are required to move simultaneously, neither has any information about the other's move before deciding their own moves.

5. Communication

Another important component influencing how people play games and how they should play is the level of communication during the game between the players. Must they make their moves without being able to speak to the other player? Would they be able to discuss with other players how they might coordinate their moves to steer the game toward an outcome more beneficial to all than would otherwise happen? Are they allowed to negotiate agreements on coordinating moves? Are the agreements binding?

At first glance, it would appear that the ability to exchange messages could never disadvantage a player. After all, he could simply choose not to talk to any of the other players. However, the ability to communicate allows one not only to offer proposals for cooperation but also the opportunity to issue threats. You might choose to say nothing, but one of the other players can announce he will play a certain way that will give you a low payoff unless you agree to play he wants the game to go. The wife in Battle of the Sexes, for example, may announce "I'm going to the prize fight whether you agree or not. If you don't come with me, then your payoff will be negative, so you'd better join me!"

6. Strategies

We come now to perhaps the central concept in all of the theory of games: strategy. By a *strategy* we mean a plan that specifies what a player should do in every possible situation that can arise in during a game. A strategy describes what move a player should make in every conceivable contingency. It is a recipe of how to play.

Consider, for example, the familiar game of tic-tac-toe (also known as Noughts and Crosses). The players are called X and O are named after the symbol each is allowed to place in an empty square of a 3×3 grid. X and O take turns. The game ends when one of the players (the winner) succeeds in placing three of his marks in a vertical, horizontal, or diagonal row of the grid. The game can end in a draw if all the squares are filled and no row, column, or diagonal is filled with the same symbol.

Suppose we number the squares in 3×3 grid as in Fig. 16.1. One strategy for playing tic-tac-toe is to place your symbol at every move in the unoccupied square with the lowest number. This strategy is not a very effective one. If my opponent becomes aware that I am using that strategy, she will simply use her first three moves to place her symbol in the

1	2	3
4	5	6
7	8	9

FIGURE 16.1 Tic-tac-toe grid.

	$C1$	$C2$	$C3$
$R1$	(r_{11}, c_{11})	(r_{12}, c_{12})	(r_{13}, c_{13})
$R2$	(r_{21}, c_{21})	(r_{22}, c_{22})	(r_{23}, c_{23})

FIGURE 16.2 A payoff matrix for a 2×3 game.

	$C1$	$C2$	$C3$
$R1$	2	3	-1
$R2$	0	-4	1

FIGURE 16.3 A payoff matrix for a zero-sum 2×3 game.

squares of the bottom row and I will lose every game. Most people learn at a fairly young age an optimal strategy for tic-tac-toe that guarantees they will at least earn a draw and can win if their opponent does not use an optimal strategy.

In many two-person games, each player has available a finite, but possibly different, number of strategies. By an $m \times n$ game, we mean a two-person game in which one player has m distinct possible strategies and the other has n distinct strategies. For an $m \times n$ game, we can construct an $m \times n$ *outcome matrix* in which each row corresponds to one of the m strategies for the first player and each column corresponds to one of the n strategies available to the second player. The entry in the ith row, jth column of the matrix, describes the end result of the game if the first player uses her ith strategy and the second player uses his jth strategy.

It's common practice in the game theory literature to use the terms *row player* and *column player*. I will follow the helpful mnemonic introduced by Peter Ungar and Philip Straffin in Straffin's book *Game Theory and Strategy:* call the first, or row, player Rose and the second, or column, player Colin. If we replace the outcome with its payoff vector, we obtain the *payoff matrix*. Fig. 16.2 shows a payoff matrix for a 2×3 game. If Rose chooses her second strategy R2 and Colin chooses his third strategy C3, then Rose's payoff is r_{23} and Colin's payoff is c_{23}.

In a zero-sum game, $c_{ij} = -r_{ij}$ for each i and j. Since Colin's payoff is the negative of Rose's for every pair of strategy choices, we know what he gets as soon as we know what she is going to receive. It is customary, then, to list only Rose's payoff in the matrix for a zero-sum game. Fig. 16.3 shows a typical payoff matrix in a zero-sum 2×3 game.

Consider the 3×4 zero-sum game with the payoff matrix shown in Fig. 16.4.

	$C1$	$C2$	$C3$
$R1$	5	3	9
$R2$	1	5	2
$R3$	2	10	6
$R4$	8	6	12

FIGURE 16.4 A payoff matrix with a dominating row.

Observe that Rose's fourth strategy gives her a higher payoff than her second strategy no matter which strategy Colin selects ($8 > 1$, $6 > 5$, $12 > 2$). In this case, we say that Rose's R4 *dominates* R2. Since R4 is universally better for Rose than R2, she would never use R2. If we eliminate row 2 from the matrix, the resulting payoff matrix becomes

$$\begin{bmatrix} 5 & 3 & 9 \\ 2 & 10 & 6 \\ 8 & 6 & 12 \end{bmatrix}$$

Noting that Colin's payoffs are the negative of Rose's, we see that Colin loses less in all cases choosing column 1 rather than column 3 so Colin would never play column 3. Eliminate the third column reduces us to the payoff matrix

$$\begin{bmatrix} 5 & 3 \\ 2 & 10 \\ 8 & 6 \end{bmatrix}$$

In this payoff matrix, Rose is always better playing in Row 3 than in Row 1, so Row 3 dominates Row 1. Cross out Row 1 to obtain the payoff matrix of the *reduced game:*

$$\begin{bmatrix} 2 & 10 \\ 8 & 6 \end{bmatrix}$$

We can provide more formal definitions about dominance. Note that one outcome is better than another outcome if it provides a higher utility to the player.

DEFINITION A strategy S *dominates* strategy T if every outcome using S is at least as good as the corresponding outcome using T and there is at least one outcome using S that is strictly better than the corresponding outcome using T. If S dominates T, then we say T *is dominated by* S and that T is a *dominated strategy*.

The *dominance principle* of game theory asserts that a rational player should never play a dominated strategy.

C. Zero-Sum Games

Consider the 3×4 two-person zero-sum game with the payoff matrix shown in Figure 16.5.

How should Rose play this game? She would certainly like the outcome R2C2, because it has the largest possible payoff for her. She could try to achieve this outcome

	*C*1	*C*2	*C*3	*C*4
*R*1	3	5	−1	−7
*R*2	1	7	2	5
*R*3	−8	1	−1	−3

FIGURE 16.5 A zero-sum 3×4 payoff matrix.

	*C*1	*C*2	*C*3	*C*4	Worst	Best of worst
*R*1	3	5	−1	−7	−7	
*R*2	1	7	[**1**]	5	1	**1**
*R*3	−8	1	−1	−3	−8	
Worst	3	7	1	5		
Best of worst			**1**			

FIGURE 16.6 Finding the saddle point of the game in Fig. 16.5.

by playing strategy R2, but she would need Colin to select C2. Colin is not likely to make that choice, as he loses 7. In fact, if Colin knew Rose was going to play R2, he would play C1 to keep his losses down to 1 unit. But if Rose knows Colin is choosing C1, she would pick R1 . . . and the reasoning continues. One of the precepts of game theory is that the players are equally intelligent and capable of equally long chains of reasoning. Each needs to keep in mind what the other would do if that player knew what we were going to do.

So Rose should be thinking that Colin is going to try his best to thwart her desire to get a large payoff. She should determine what is her worst possible outcome for each of her strategies. For R1, it is −7; for R2, it is 1; and for R3 it is −8. The "best of the worst" is 1, so if she plays R2, Rose is guaranteed to get a payoff of at least 1. There is nothing Colin can do to force a smaller payoff to Rose. The "best of worst" criterion is called the *maximin* approach; Rose first determines the *minimum* in each row and then chooses the *maximum* of these numbers. We call this number the *lower value* of the game.

Colin's considerations are similar to Rose's. He determines his worst possible outcome for each of his strategies. He does this by finding the largest possible number in each column because that would be the magnitude of his loss. Thus, he looks for the *maximum* in each column. Of these maxima, he wants to choose the smallest one, the *minimum*. For the column player, the "best of the worst" is the *minimax*. This number constitutes the *upper value* of the game. Colin can guarantee that Rose will never get more than this amount no matter what scheme she follows.

For this particular game, the lower value and the upper value coincide. The common number is called a *saddle point* of the game. It is simultaneously the smallest number in its row and the largest number in its column. The strategy choices whose outcomes intersect at this payoff are optimal strategies for the players in a zero-sum game. The row player can guarantee herself that she will get at least the lower value and the column player can assure himself that his losses will be no more than the upper value. When the values are equal, neither player can do any better. In the game of Fig. 16.6, Rose should play R2 and Coin should play C3.

	$C1$	$C2$	Worst	Lower Value
$R1$	2	10	2	
$R2$	8	6	6	**6**
Worst	8	10		
Upper Value	**8**			

FIGURE 16.7 A 2×2 zero-sum game with no saddle point.

Outcome	**Probability**	**Value**
R1C1	Pq	2
R1C2	$p(1-q)$	10
R2C1	$(1-p)q$	8
R2C2	$(1-p)\,(1\,-q)$	6

FIGURE 16.8 The outcomes of the game in Fig. 16.7 with probability and payoffs.

The *saddle point principle* of game theory asserts that if a game has a saddle point, then rational players should play a strategy that contains a saddle point.

In general, however, the lower value of a game is usually strictly smaller than the upper value and there is no saddle point. The reduced game of Fig. 16.7 is a simple example. Here the lower value is 6 and the upper value is 8. See Fig. 16.7. Rose can get at least 6 and Colin can hold her payoffs down to 8. Is there some way to play this game so that Rose gets more 6? Can we attach a single value to this game?

The answers to both questions are *No* if we constrain ourselves to so-called "pure" strategies. Game theory introduces the idea of a *mixed strategy* to get *Yes* answers. Imagine what might happen if Rose and Colin play this game repeatedly where the players chose their strategies with different frequencies, sometimes using the first strategy and other times the second? Suppose Rose chooses at random between R1 and R2, selecting R1 with probability p and R2 with probability $1 - p$ while Colin is randomly picking between his two strategies, employing C1 with probability q and C2 with probability $1 - q$. It is natural then to examine the *expected value of the payoff*.

Recall that the expected value is the weighted sum of all possible outcomes, each weighted by its probabilities of occurring. Since Rose and Colin make their choices independently of each other, we can combine probabilities by multiplying them.

Rose's expected value is $\mathrm{EV}_{\mathrm{Rose}} = 2pq + 10p(1-q) + 8(1-p)q + 6(1-p)(1-q)$. If we multiply out and collect like terms, we find

$$\mathrm{EV}_{\mathrm{Rose}} = 6 - 10pq + 4p + 2q$$

which we may write as

$$\mathrm{EV}_{\mathrm{Rose}} = 6\frac{4}{5} - 10\left(p - \frac{1}{5}\right)\left(q - \frac{2}{5}\right)$$

In this form, we see that if Rose chooses $p = 1/5$, then her expected payoff will be 6.8 regardless of Colin's choice of q. If Colin sets $q = 2/5$, then he holds Rose's payoff to 6.8 for

every choice of p. What if Rose tries to increase her expected payoff? She would need to make the term $-10\left(p-\frac{1}{5}\right)\left(q-\frac{2}{5}\right)$ positive. If she choose $p<1/5$, then $-10\left(p-\frac{1}{5}\right)$ becomes positive, but Colin can respond with a q less than 2/5, rendering $-10\left(p-\frac{1}{5}\right)\left(q-\frac{2}{5}\right)$ negative and decreasing Rose's expected payoff to something below 6.8. On the other hand, if Rose tries a p larger than 1/5, Colin can answer with $q>2/5$, again lowering Rose's expected payoff below 6.8. Thus, Rose's attempt to be greedy can backfire. Similarly, if Colin tries on his own to decrease his losses by making Rose's expected payoff smaller than 6.8 by tinkering with his q, she can find an appropriate p that will *increase* her expected payoff. So neither player can benefit from moving away for choosing $p=1/5$, $q=2/5$; in fact, each runs the risk of doing worse. We have the equivalent of a saddle point in mixed strategies. In this game there are optimal mixed strategies for each player. The optimal mixed strategies provide an equilibrium solution for the game. The value of the game is the expected payoff to Rose if she and Colin use the optimal mixed strategies.

How does Rose implement a mixed strategy of playing R1 one-fifth of the time and R2 four-fifths? One way is to use R1 in the first game, followed by R2 in the next four games and then repeat this pattern over and over again. But Rose is in trouble if Colin figures out the pattern. Colin can then play C1 every time Rose plays R1 and C2 each time she uses R2. Then Rose gets 2 with frequency 1/5 and 6 with frequency 4/5, yielding an average payoff of only 5.2. The result is similar if Rose follows any systematic pattern for choosing R1 and R2. If Colin discovers the pattern, he can find a matching one of his own to depress her expected payoff. Rose must avoid using a pattern. She should use a randomizing device to make her choice. She could construct a simple spinner, for example, with 4/5 of the wheel colored red and 1/5 colored blue. Before each game, she spins the needle. If it lands on blue, play R1; otherwise, play R2. [Rose might also make use of a pseudorandom number generator of the type discussed in Chapter 15.] Since Rose does not know in advance which of R1 or R2 she is going to use in the next play of the game, Colin has no way to anticipate her choice and exploit a pattern.

To summarize the results of our analysis of the game of Fig. 16.6: Each player in this 2×2 zero-sum game has an optimal mixed strategy that he or she should use in a randomized fashion. The value of the game is the expected payoff under these optimal mixed strategies.

The first major theorem in game theory generalizes this result to an arbitrary $m\times n$ two-person zero-sum game. Called the Minimax Theorem, it was first proved by John von Neumann in 1928. Von Neumann showed that there would always exist a pair of optimal mixed strategies in a zero-sum game between two players where each had a finite number of pure strategies. The expected value v using the optimal mixtures is a number the row player can guarantee receiving and the column player can guarantee as a bound on his loss. If either player deviates from the optimal mixture, the opposing player has a mixed strategy response, which will punish the deviating player.

There are many different proofs of von Neumann's Minimax Theorem. In Section IV of this chapter we will outline a proof of a generalization of this theorem due to John Nash that makes use of the Brouwer Fixed Point Theorem.

III. The Binding of Isaac

Let's return to the story of Abraham and the possible sacrifice of Isaac and examine it through the lens of game theory. We shall follow the analysis of Steven Brams, who pioneered in the application of game theory to Bible narratives. The outcome of this

situation depends on the decisions of both Abraham and God. Abraham must decide whether to go forward with the sacrifice and God must decide how to react to whatever choice Abraham makes. Game theory aims, in part, to determine what strategy rational players should follow to best achieve their objectives. Is it presumptuous to treat God as a mere player in a game? Isn't the Almighty ultimately unknowable and not subject to our human understanding? Brams gives a carefully reasoned response:

> *In any biblical analysis . . . God must be given His proper due. He is* the central character in the Bible. Accordingly, I propose to treat Him as such, but my treatment assumes more than His omnipresence. I also assume that God is motivated to do certain things—that He has goals He would like to achieve.
>
> *I do not assume that God is omnipotent . . . the Bible is clear on one thing: human beings do* have free will and can exercise it, even if it invokes God's wrath. . . . Consequently, God, powerful as He is, is sometimes thwarted.
>
> *Since God does not always get His way, He can properly be viewed as a participant, or player*, in a game.

Additional evidence to support Brams's view comes from earlier sections of the Bible, which recount Abraham's interactions with God. We learn that Abraham can appeal to God's preferences to persuade Him to change His mind. In Chapter 18 of Genesis, for example, God informs Abraham of His intent to destroy the sinful cities of Sodom and Gomorrah. Abraham challenges God, in an almost rebuking voice "Will you sweep away the innocent along with the guilty? What if there should be fifty innocent within the city; will You then wipe out the place and not forgive it for the sake of the innocent fifty who are in it? Far be it from You to do such a thing, to bring death upon the innocent as well as the guilty, so that innocent and guilty fare alike. Far be it from You! Shall not the Judge of all the earth deal justly?"

God agrees to spare Sodom and Gomorrah if there are 50 innocent among the people. Abraham is not content to accept this judgment. Outwardly humble in speech ("I who am but dust and ashes"), he demands to know what God would do "if the fifty innocent should lack five? Will you destroy the whole city for want of the five." God relents and agrees to spare Sodom and Gomorrah if there are 45 innocent people. In the next several verses, Abraham gets God to lower the number even further, first to 40, then to 30, then 20, and finally to 10.

Let's see what happens then if we regard both God and Abraham as players in a two-person game. Abraham has two choices: **O**: he can offer Isaac as a sacrifice or **O***: he refuses to go forward with the sacrifice. God also has two possible moves: He can act with mercy (**R**) or He can be adamant (**R***).

Fig. 16.9 shows the outcome matrix for the game. There are four scenarios:

OR: Abraham offers Isaac as a sacrifice and God is merciful.

OR*: Abraham offers Isaac as a sacrifice and God is adamant.

O*R: Abraham doesn't sacrifice Isaac and God is merciful.

O*R*: Abraham doesn't sacrifice Isaac and God is adamant.

How do we go from an outcome matrix to a payoff matrix in this case? We do not need to know the utility functions of Abraham and God to do the analysis. It will be sufficient to

		GOD	
		R	**R***
	A	Abraham faithful	Abraham faithful
	B **O**	God merciful	God Adamant
	R	Isaac saved	Isaac sacrificed
	A		
	H	Abraham resistant	Abraham resistant
	A **O***	God merciful	God Adamant
	M	Isaac saved	Isaac's fate uncertain

FIGURE 16.9 The outcome matrix for the Abraham-God game.

	R	R^*
O	4	3
O*	1	2

FIGURE 16.10 How God ranks the outcomes.

	R	R^*
O	4	3
O*	2	1

FIGURE 16.11 How supremely faithful Abraham ranks the outcomes.

rank order their preferences. We will use the numerals 1, 2, 3, and 4, with a higher number indicating a more preferred outcome. Thus, we label the most preferred outcome with a 4 and the least preferred with a 1. We are only paying attention to order here. We are not asserting that the outcome with a 4 is twice as desirable as an outcome with a 2, for example.

We assume that God would prefer to be obeyed than disobeyed. Thus, the entries in the first row of the matrix will get the 4 and 3 designations, while 1 and 2 will be assigned to the second row entries. If Abraham obeys God's command, so that we are in the top row of the matrix, it is reasonable to assume that God would prefer to be merciful; Abraham will have passed God's test, so why punish him? On the other hand, should Abraham disobey, then God will not be happy and would prefer to be adamant rather than merciful. Thus, we have God's preference order OR > OR* > O*R* > O*R, which we display in matrix form in Fig. 16.10.

What are Abraham's preferences among the possible outcomes? We can identify three different views of Abraham. First, there is the Abraham who is faithful to God whatever the circumstances are. For such an Abraham, showing his faith is paramount; he would rather obey than disobey. Thus, his two most desired outcomes (4 and 3) will be in the top row of the payoff matrix and the two least preferred (2 and 1) will be in the bottom row. Certainly Abraham would prefer that God act mercifully rather than adamantly; he prefers outcomes in the left column over outcomes in the right column. Thus, Abraham ordering, from most to least preferred, would be OR > OR* > O*R > O*R*. Fig. 16.11 shows the ever faithful Abraham's payoffs.

$$\begin{array}{c} \\ \mathrm{O} \\ \mathrm{O}^* \end{array} \begin{array}{c} \begin{array}{cc} R & R^* \end{array} \\ \begin{pmatrix} 4 & 2 \\ 3 & 1 \end{pmatrix} \end{array}$$

FIGURE 16.12 How a somewhat wavering Abraham ranks the outcomes.

$$\begin{array}{c} \\ \mathrm{O} \\ \mathrm{O}^* \end{array} \begin{array}{c} \begin{array}{cc} R & R^* \end{array} \\ \begin{pmatrix} 4 & 1 \\ 3 & 2 \end{pmatrix} \end{array}$$

FIGURE 16.13 How a seriously wavering Abraham ranks the outcomes.

The second view of Abraham shows his faith wavering somewhat. He would most prefer that God act with mercy rather than adamantly, so 4 and 3 will appear in the left column with 2 and 1 in the second column. It is still very important to Abraham that he display his faith, so he would rather obey than disobey. This Abraham has the preference order OR > O*R > OR* > O*R*. Fig. 16.12 shows Abraham's payoffs with these assumptions.

A third perspective on Abraham is that his paramount concern is not showing faith in God, but rather in saving Isaac's life. This Abraham has a preference ordering similar to the previous case, but if God were going to be adamant, Abraham would rather not carry out the sacrifice. The outcome OR* would be the worst possible for Abraham for in this case it is certain that Isaac will die. If Abraham withholds Isaac (O*) and God is adamant so O*R* results, then he expects some punishment from God, but it may be entirely directed at Abraham, with Isaac being spared. Of course, if God were going to be merciful, Abraham would prefer to obey. The preference ordering for this Abraham among the four possible outcomes would be OR > O*R > O*R* > OR*. See Fig. 16.13.

Let's turn to analyzing the three games that correspond to this trio of views about Abraham. What is common in all three is that Abraham makes the first move and then God responds. Abraham has only two strategies: obey (O) or disobey (O*). God, however, has four strategies, since He has two choices (R or R*) for each of Abraham's moves. We will use the notation A/B to describe a strategy for God where God chooses A if Abraham obeys and chooses B if Abraham disobeys. We can denote and describe God's possible strategies as follows:

R/R: Be merciful regardless. If Abraham offers Isaac, renege on your command, and intervene to stop the sacrifice. If Abraham refuses to sacrifice Isaac, relent; show mercy and do not punish him.

R*/R*: Be adamant regardless. If Abraham offers Isaac, let the sacrifice be completed. If Abraham withholds his son, punish him for disobedience.

R/R*: Tit-for-Tat: Show mercy and stop the sacrifice if Abraham is obeying God, but punish him if he won't kill Isaac.

R*/R: Tat-for-Tat: Be adamant if Abraham obeys and be merciful if he doesn't. An adamant God would let the sacrifice continue to its bitter end should Abraham display his obedience. Acting mercifully if Abraham disobeys would mean forgiving him.

FIGURE 16.14 Payoff matrix for Case (a): Abraham faithful regardless.

	R	R^*	R/R	R^*/R^*	R/R^*	R^*/R
O	(4, 4)	(3, 3)	(4, 4)	(3, 3)	(4, 4)	(3, 3)
O*	(2, 1)	(1, 2)	(2, 1)	(1, 2)	(1, 2)	(2, 1)

Fig. 16.15a Payoff matrix for Case (b): Abraham wavers somewhat.

	R	R^*	R/R	R^*/R^*	R/R^*	R^*/R
O	(4, 4)	(2, 3)	(4, 4)	(2, 3)	(4, 4)	(2, 3)
O*	(3, 1)	(1, 2)	(3, 1)	(1, 2)	(1, 2)	(3, 1)

Fig. 16.15b Payoff matrix for Case (c): Abraham wavers seriously.

	R	R^*	R/R	R^*/R^*	R/R^*	R^*/R
O	(4, 4)	(1, 3)	(4, 4)	(1, 3)	(4, 4)	(1, 3)
O*	(3, 1)	(2, 2)	(3, 1)	(2, 2)	(2, 2)	(3, 1)

For each of our three possible Abrahams, we have a 2×4 game. For each we will display an additional two columns on the left showing the payoffs for the four possible outcomes of the game. Fig. 16.14 shows the result for a supremely faithful Abraham.

In this payoff matrix, Abraham always does better in the first row than in the second regardless of which strategy God chooses: $4 > 2$ (R/R), $3 > 1$ (R^*/R^*), $4 > 1$ (R/R^*), and $3 > 2$ (R^*/R). Thus, Abraham's obey strategy is dominant and he will always choose it. God recognizes this as well, so He can secure His best outcome (4) by choosing R/R or R/R*. For either choice, the outcome of the game is that Abraham will obey God and proceed with the sacrifice, but God will be merciful and intervene. The outcome is OR; both players get their best possible payoff.

The payoff matrix for an Abraham who wavers somewhat in his faith appears in Fig. 16.15.

Here Abraham does not have a dominant strategy. Obeying (O) is better against R/R, R*/R*, and R/R*, but O* is better if God plays R*/R. Abraham needs to think about what God will do before he can decide what strategy he should follow. Viewing the payoff matrix from God's perspective, we see that R/R*, the tit-for-tat strategy, is a dominant one for God. It is always better than R*/R: $4 > 3$ and $2 > 1$. It is at least as good as R/R in all cases and better in one case: $4 \geq 4$ (if Abraham chooses O) and $2 > 1$ (if Abraham chooses O*). Similarly, R*/R is better for God than R^*/R^* $(4 > 3)$ if Abraham obeys and at least as good for God $(2 \geq 2)$ if Abraham disobeys. Thus, R/R* is a dominant strategy for God. God will choose R/R*. Abraham, realizing this, knows he will get 4 if he plays O and 1 if he chooses O*. Thus, Abraham will choose the O strategy. God, playing the tit-for-tat strategy, will be merciful. The outcome is again OR and each gets the best possible payoff.

Finally, let's do the analysis for our seriously wavering Abraham for whom saving Isaac is more important than showing his faith. Here the payoff matrix for the 2×4 game is shown in Fig. 16.15b.

As in the previous case, Abraham does not have a dominant strategy. O is better against R/R and R/R*, but O* is better against R*/R* and R*/R. Abraham is forced to look at what God might do. Abraham sees that R/R* is again the dominant strategy for God. Knowing that God will choose R/R*, Abraham will choose 0. The outcome, once more, is OR and yet again both players get their best possible payoff.

The conclusion for game theory is that if the players act rationally and choose dominating strategies when they have them, then no matter which of the three views of Abraham we have, the outcome will be the same. Abraham will attempt to carry out God's order to sacrifice his son, but God will intervene to stop it.

How accurate is game theory's prediction? Let's go back to Genesis for the conclusion of the story. Recall that we interrupted the narrative just after Abraham had bound Isaac and laid him on altar: "Then he reached out his hand and took the knife to slay his son."

The text continues:

> *But the angel of the Lord called out to him from heaven, "Abraham! Abraham!"*
>
> *"Here I am," he replied.*
>
> *"Do not lay a hand on the boy," he said. "Do not do anything to him. Now I know that you fear God, because you have not withheld from me your son, your only son."*
>
> *Abraham looked up and there in a thicket he saw a ram caught by its horns. He went over and took the ram and sacrificed it as a burnt offering instead of his son. So Abraham called that place The Lord Will Provide. And to this day it is said, "On the mountain of the Lord it will be provided."*
>
> *The angel of the Lord called to Abraham from heaven a second time and said, "I swear by myself, declares the Lord, that because you have done this and have not withheld your son, your only son, I will surely bless you and make your descendants as numerous as the stars in the sky and as the sand on the seashore. Your descendants will take possession of the cities of their enemies, and through your offspring all nations on earth will be blessed because you have obeyed me".*

Thus, the outcome game theory predicted is how the story turned out. Does game theory provide any further insight into the Akedah that classic biblical commentary and literary analysis fails to give? It is certainly consistent with the game theory approach that Abraham's action followed the traditional interpretation that he passed God's test by demonstrating unwavering faith in blindly obeying the command to sacrifice Isaac. But game theory says there is another possibility: instead of acting out of faith, Abraham might have had other priorities. His decision to move forward with the sacrifice could be adequately explained as the logical action to take in light of a dispassionate, cold, rational analysis of the game. A game theorist could well conclude that if God's test was designed to be a test of faith, it was a poor exam. Abraham could have passed the test even if his faith was quite weak.

IV. Tosca and the Prisoners' Dilemma

The *Prisoners' Dilemma* is justifiably one of the most studied games in the history of mathematics. Merrill Flood and Melvin Dresher created the game in 1950 as an example of a situation where people might not cooperate even if it is in their best interests to do so.

Albert Tucker (1905−1995) framed the example in its current form when he wanted to come up with an interesting example to explain game theory to a class of psychology students at Stanford a half-century ago.

Police suspect that a certain pair of men, Ehrlichperson and Handlebody, have committed a major crime together, a felony punishable by 20 years in prison. They are arrested, jailed, and interrogated separately. There is sufficient evidence to charge them with a minor offense, but without a confession, the state will not be able to convict them of the more serious crime. The prisoners are aware that there is enough evidence in police hands to send them to jail for 1 year.

The district attorney offers each man the same deal. "If you confess to the major crime and implicate your partner, then he will receive a 20-year prison sentence, but you will go free for being a state's witness. On the other hand, if you remain silent and he confesses that you did the crime together, then you will be the one with the 20-year prison term, and he will be free. If you both confess, then we will let the pair of you enter into a plea bargaining deal under which you both serve 5-year terms." If both prisoners remain silent, they will each spend a year behind bars.

Each player has two strategies: Confess (C) or Remain Silent (S). Fig. 16.16 shows the payoff matrix of the classic Prisoners' Dilemma Game:

What should each of the prisoners do? They are not able to communicate with each other, so each must make a decision without knowing for certain what the other prisoner will do. This seems easy because each has a dominant strategy: Confess. For Ehrlichperson, for example, C is a better choice if Handlebody chooses C; he would be in prison for 5 years instead of 20. C is also better for Ehrlichperson if Handlebody remains silent; he goes free if he confesses but spends a year in prison if he remains silent. Handlebody has a similar analysis; he is better off confessing no matter what Ehrlichperson does.

Thus, maximizing your own gains (or equivalently minimizing your individual losses) by choosing a dominant strategy if you have one leads to the outcome of both choosing strategy C. Each faces 5 years locked in a penitentiary.

But wait! There is another strategy choice that leads to a better outcome for both of them! If each chooses to remain silent (S), then each only goes to prison for a year. Ehrlichperson and Handlebody both prefer $(-1, -1)$ to $(-5, -5)$. The Confess (C) is often described as an act of betrayal because it results in the other player getting a longer prison sentence than he would have if the confessor had remained silent. Betrayal is a dominant strategy in the Prisoners' Dilemma, and the Dominance Principle says that players should always choose a dominating strategy. Following that advice, however, leads to a less desirable outcome for both than if each had remained silent.

		Handlebody	
		Confess C	*Stay Silent S*
Ehrlichperson	*Confess C*	$(-5, -5)$	$(0, -20)$
	Stay Silent S	$(-20, 0)$	$(-1, -1)$

FIGURE 16.16 Payoff matrix for prisoners' dilemma.

The arms race between the United States and the Soviet Union during the Cold War exhibited some characteristics of the Prisoners' Dilemma. Each nation had a choice to arm or to disarm. Disarming while your opponent continued to build up arms was seen as leading to military inferiority (the "missile gap") and possible annihilation. On the other, if you arm and the other disarms, you have military superiority. If each side had huge stockpiles of nuclear weapons, neither could afford to attack the other without fear of its own destruction, but both sides would have a large economic burden that would deflect expenditures away from important domestic concerns. If both sides chose to disarm, then neither would be in danger from the other and each would free up revenues for other purposes. The best choice is mutual disarmament, but the "rational" choice appeared to be a mutual escalation in weaponry and that's what both countries did for several decades.

It turns out that the situation posed in Puccini's opera is a Prisoners' Dilemma for Tosca and Scarpia. Each one has two strategies: keep the bargain made with the other one or double cross the other. Scarpia's double cross is signaling his assistant that the "fake" execution should be a real one. Tosca's double cross is to kill Scarpia. Fig. 16.17 displays the outcome matrix for the game.

For Tosca, the best outcome occurs if she double-crosses Scarpia but he keeps his bargain. She is reunited with her beloved Cavaradossi and avoids sex with the now dead Scarpia, who will never bother her again. The second-best outcome is the one in which Cavaradossi lives, but she must sully herself with Scarpia's lust. The worst outcomes for Tosca are the two in which Cavaradossi dies; of these, the worst is the one where she keeps her bargain but Scarpia has double-crossed her. Using the numerals 1 to 4 as in the Abraham story, we assign them to the outcomes as we just described. Fig. 16.18 shows the payoffs to Tosca, Fig. 16.19 for Scarpia, and Fig. 16.20 combines them for a full payoff matrix.

Scarpia		**Keep Bargain**	**Double Cross**
Tosca	**Keep Bargain**	Scarpia's lust is satisfied Cavaradossi lives Tosca's virtue is compromised	Scarpia's lust is satisfied Cavaradossi dies Tosca's virtue is compromised
	Double Cross	Scarpia dies Cavaradossi lives Tosca's virtue remains intact	Scarpia dies Cavaradossi dies Tosca's virtue remains intact

FIGURE 16.17 Outcome matrix for Tosca and Scarpia.

Scarpia		**Keep Bargain**	**Double Cross**
Tosca	**Keep Bargain**	3	1
	Double Cross	4	2

FIGURE 16.18 Payoffs to Tosca.

		Scarpia	
		Keep Bargain	**Double Cross**
Tosca	**Keep Bargain**	3	4
	Double Cross	1	2

FIGURE 16.19 Payoffs to Scarpia.

		Scarpia	
		Keep Bargain	**Double Cross**
Tosca	**Keep Bargain**	(3, 3)	(1, 4)
	Double Cross	(4, 1)	(2, 2)

FIGURE 16.20 The payoff matrix for the game in *Tosca*.

It is easy to see from the payoff matrix of Fig. 16.20 that the dominant strategy for both Tosca and Scarpia is to double-cross each other. The outcome is that each gets their second-worst payoff. If they both choose the keep the bargain, then each winds up with their second-best outcome. Did they find that better outcome, or does the opera conclude with a double double cross, or perhaps some other outcome?

As Scarpia sings "Tosca, you are mine at last!" he opens his arms and advances towards Tosca to embrace her . . .

The libretto describes what happens next:

> *Scarpia's shout of lust ends in a cry of anguish. Tosca has struck him full in the heart with the knife. "Accursed one!" exclaims Scarpia to which Tosca triumphantly answers "This is the kiss of Tosca!"*

Both our players went for the dominant strategy. Tosca decided to deceive Scarpia by appearing to agree to his demand, but then stabbing him dead after he has given the order to use blanks. She does so, but too late discovers that Scarpia chose a double cross as well. The firing squad does not use blanks; Cavaradossi dies. Tosca leaps from the battlements, committing suicide, and all three end up dead.

The Prisoners' Dilemma Game illustrates one essential difference between nonzero-sum games and zero-sum situations. For zero-sum games, we only have to consider our own payoffs is selecting strategies that are most likely to give us the best possible outcomes. For nonzero-sum games, *rational* no longer can means maximizing our expected value, thinking the worst about the other player. Such selfish thinking can lead to the paradoxical outcome that members of a group will consciously steer towards a sub-optimal outcome in certain scenarios. To do as well as possible in a nonzero-sum game, the theory needs to take into account, in advising us of our most attractive strategy mixture, other players' payoffs as well as our own.

V. Nash Equilibrium

The mutual double cross outcome in the Tosca-Scarpia game (see Fig. 16.20) has an equilibrium aspect to it. If either player moves away from this outcome acting individually, then the payoff to the one who moves decreases. We are currently at (2, 2). If Tosca moves to her Keep Bargain strategy, the payoffs shift to (1, 4) where Tosca winds up with her worst possible outcome. If Scarpia moves alone to his Keep Bargain strategy, the new payoff is (4, 1); his payoff decreases from 2 to 1. Thus, neither player has an incentive for a unilateral move to another strategy. They do have a mutual incentive to move to the (3, 3) payoffs, but they may not be able to reach an agreement to do so either because they are not able to communicate with each other or there isn't sufficient trust of one another.

In his short, but brilliant Ph.D. thesis, John F. Nash generalized this situation and developed what is now called the *Nash equilibrium*. The setting is a game with at least two players who act independently of each other without the ability to make binding agreements. The term *noncooperative game* is often employed to describe such games. Each player can see the payoff matrix. A Nash equilibrium is an assignment of strategy choices to the players so that no player can benefit by changing strategies if the other players keep theirs unchanged.

As we observed, the Double Double Cross is a Nash equilibrium for Tosca and Scarpia. The outcome in which they both choose Keep Bargain and receive (3, 3) is not a Nash equilibrium. While Tosca's payoff would go down if she switches to Double Cross and Scarpia doesn't move, Scarpia's payoff would go up if he switched to Double Cross while Tosca continued to use the Keep Bargain strategy.

A strategy combination (R_i, C_j) in a two-person game is a Nash equilibrium pair if no player can increase his or her reward by a unilateral departure from $(R_i\ C_j)$. If one player sticks rigidly to his or her Nash-equilibrium strategy, then the other player cannot increase his or her payoff by selecting a strategy other than his or her Nash-equilibrium strategy. The pair (R_i, C_j) is a Nash equilibrium if R_i is the best reply to C_j and C_j is the best reply to R_i

As another example, consider the payoff matrix shown in Fig. 16.21 for a 2×2 two-person game. Both R1C2 and R2C1 produce Nash equilibria. If the players find themselves at R1C2, for example, then neither Rose nor Colin has any incentive to move on his or her own. Rose sees that her move to R2 results in the outcome R2C2 where payoff drops from 10 to 5. Colin realizes similarly that if he moves and Rose doesn't, his payoff drops. The new outcome would be R1C1 where Colin's payoff now becomes -10 instead of the original -1.

Game theorists use the name *chicken* for games that have payoffs with the structure shown in Fig. 16.21. Such games may occur in situations where each player would rather not yield to the other, but the worst possible outcome happens if neither yields. The name *chicken* refers to a contest where two drivers speed toward each other on a collision course.

	$C1$	$C2$
$R1$	$(-10, -10)$	$(10, -1)$
$R2$	$(-1, 10)$	$(5, 5)$

FIGURE 16.21 The game of chicken.

They will both die if neither one swerves out of the way of the other, but the driver who swerves is labeled a coward or "chicken." This game has also used a model of the nuclear brinksmanship involved in the Cuban Missile Crisis.

Biologists know this game as *Hawk-Dove* where two animals are contesting an indivisible resource. If both use the more aggressive Hawk strategy, then they fight until one is injured and the other wins. If one chooses Hawk and the other the less aggressive Dove strategy, then Hawk beats Dove. If both employ Dove, there is a tie; each receives a payoff smaller than the payoff a Hawk gets in beating a Dove. See Maynard Smith [1982] and the discussion in Chapter 18. Chicken has also been used to model some economic decisions. Suppose two companies are considering entering a market in which there is a relatively low demand for a product so that there is only enough room in the market place for one of them. If both enter, then each will go broke. See Krugman [1987].

Some games such as Tosca-Scarpia have a single Nash equilibrium, whereas others, such as chicken, may have multiple Nash equilibria. A more common situation is that in which no pair of strategies produces a Nash equilibrium. Fig. 16.22 shows such a game. Note that Rose does better by a unilateral move from R2C1 or R1C2, while Colin improves his payoff by a unilateral move from R1C1 or R2C2. None of the four possible outcomes is a Nash equilibrium.

In the case of zero-sum games with no saddle points in pure strategies, we proceeded to consider mixed strategies where players chose each pure strategy with a probability picked in order to maximize expected payoffs. We can also examine Nash equilibria with mixed strategies.

In the Rose-Colin game with the payoff matrix shown in Fig. 16.22, suppose Rose plays strategy R1 with probability p and strategy R2 with probability $(1-p)$. Let's determine how well Colin will do.

Colin's expected payoff if he always chooses C1 is

$$-3p+5(1-p)=-3p+5-5p=5-8p$$

and his expected payoff under C2 is

$$4p-4(1-p)=4p-4+4p=8p-4.$$

Then C1 will have a higher expected payoff than C2 if $5-8p>8p-4$—that is, when $p<\frac{9}{16}$. Colin's second strategy C2 has a higher expected payoff when $p>\frac{9}{16}$. The two strategies have the same expected payoff, $\frac{1}{2}$, when $p=\frac{9}{16}$. Thus, if Rose uses R1 with probability $\frac{9}{16}$ and R2 with probability $\frac{7}{16}$, then Colin will have an expected payoff of $\frac{1}{2}$ for every mixture of C1 and C2 that he chooses.

	$C1$	$C2$
$R1$	$(5,-3)$	$(-4,4)$
$R2$	$(-5,5)$	$(3,-4)$

FIGURE 16.22 A 2×2 game with no pure Nash equilibrium.

Shifting perspectives, we find that if Colin uses a mixture where he plays C1 with probability q and C2 with probability $(1-q)$, then Rose's expected payoffs are $(5q + -4(1-q) = 5q - 4 + 4q = -4 + 9q)$ if she always plays R1 and $(-5q + 3(1-q) = -5q + 3 - 3q = -8q + 3)$ if she always plays R2. These expected payoffs are equal when $q = \frac{7}{17}$ and have a value $\frac{-5}{17}$. If Colin uses the strategy mixture $\left(\frac{7}{17}, \frac{10}{17}\right)$, then Rose's expected payoff would be $\frac{-5}{17}$, no matter what strategy mixture she picked.

Our claim is that the strategy mixtures $\left(\frac{9}{16}, \frac{7}{16}\right)$ for Rose and $\left(\frac{7}{17}, \frac{10}{17}\right)$ for Colin provide a Nash equilibrium for the game. If either player sticks to the suggested mixture and the other deviates to another mixture, the deviating individual will see no change at all in the expected payoff. There is no mixture that will increase his or her expected payoff.

For this particular two-person game, we were able to compute mixed strategies that yielded a Nash equilibrium. Nash proved a major extension of this result: namely, there always exists at least one Nash equilibrium in any n-person game. We will now outline a proof of Nash's theorem.

First, we need to introduce some notation: we let n be the number of players and we use i, j, k as indices for individual players. Lowercase Greek letters such as α, β, γ denote indices for the set of strategies available to an individual. The symbol $\pi_{i\alpha}$ means the αth pure strategy of individual i.

By a *mixed strategy s_i for player i* we mean a collection $\{c_{i\alpha}\}$ of nonnegative numbers that sum to 1 and are in a one-to-one correspondence with player i's pure strategies. Our symbol for such a mixed strategy is $s_i = \sum_\alpha c_{i\alpha}\pi_{i\alpha}$ where each $c_{ia} \geq 0$ and $\sum_\alpha c_{i\alpha} = 1$. The number c_{ia} is the probability that player i will use his αth strategy. In addition to s_i as a symbol for a mixed strategy, we will also use terms like t_j and r_k. Since each of our n players can select a mixed strategy that may differ from one individual to another, we can consider n-tuples $S = (s_1, s_2, \ldots, s_n)$ of mixed strategies. For each collection S of mixed strategies, there will be an n-dimensional vector $\mathbf{p}$ of expected payoffs. We call $\mathbf{p}$ the *payoff function*. Then $\mathbf{p}_i(S) = \mathbf{p}_i(s_1, s_2, \ldots, s_n)$ indicates the expected payoff to player i.

Given a particular set S of strategy mixtures, we need a notation for a new set of mixtures in which exactly one of the players switches strategies. Let $(S;t_i)$ indicate the new set where t_i replaces the original s_i for player i. More exactly, if $S = (s_1.s_2,s_{i-1},s_i,s_{i+1} \ldots,s_n)$ then (S, t_1) $(s_1.s_2,s_{i-1},t_i,s_{i+1} \ldots,s_n)$

With this notation, we can provide a precise description of a Nash equilibrium. An n-tuple S is a Nash *equilibrium point* if and only if for every player i,

$$p_i(S) = \max_{\text{all r}'_i\text{ s}} [p_i(S; r_i)]$$

This is a precise way of saying that each player's mixed strategy in S maximizes his payoff if the strategies of the others are held fixed. What is true for every player i is that no matter what other mixture r_i he tries, his payoff will not increase if he is the only one changing strategy mixtures.

It's easy to show that $\max_{\text{all r}'_i\text{ s}} [p_i(S; r_i)] = \max_\alpha [p_i(S; \pi_{i\alpha})]$. This result implies that if in replacing a mixed strategy by any of the pure strategies a player has does not increase her expected payoff, then no mixture will either. Thus, to test whether a particular set of mixed strategies is a Nash equilibrium, we do not have to examine the infinitely many

possible other mixtures for each of the players, but only the finite set of pure strategies for each player. If player i replaces her strategy mixture s_i by her pure αth strategy, then we denote her expected payoff $p_i(S;\pi_{i\alpha})$ by $p_{i\alpha}(S)$.

Each of our players typically has a large number of pure strategies available. A particular strategy mixture may attach 0 as the weight of one or more pure strategies so that the corresponding pure strategy is never employed. If the weight $c_{i\alpha}$ is positive, however, then we say that mixture *uses* the pure strategy $\pi_{i\alpha}$.

A necessary and sufficient condition for S to be a Nash equilibrium is

$$p_i(S) = \max_{\text{all } r'_i \text{ s}} [p_i(S;r_i)] = \max_{\alpha} [p_i(S;\pi_{i\alpha})] = \max_{\alpha} p_{i\alpha}(S)$$

Note: if S is a Nash equilibrium point, then $c_{i\alpha=0}$ whenever $p_{i\alpha}(S) < \max_\beta p_{i\beta}(S)$—that is, S does not use $\pi_{i\alpha}$ unless it is an optimal pure strategy for player i.

Hence: A necessary and sufficient condition for a Nash equilibrium is

If $\pi_{i\alpha}$ is used in S, then $p_{i\alpha}(S) = \max_\beta p_{i\beta}(S) p_{i\alpha}(S) = \max_\beta p_{i\beta}(S)$.

We come now to a proof of Nash's theorem guaranteeing the existence of at least one Nash equilibrium. The proof is very similar in spirit and structure to the argument for the existence of a price equilibrium that we introduced in Chapter 9. We suggest you review that argument before continuing.

NASH'S THEOREM: Every finite game has an equilibrium point.

Proof Let S be an n-tuple of mixed strategies and $p_{i\alpha}(S)$ the corresponding payoff to player i if he changes to his αth pure strategy $\pi_{i\alpha}$ and the others continue to use their respective mixed strategies from S. Define $\varphi_{i\alpha}(S) = \max(0, p_{i\alpha}(S) - p_i(S))$, and for each component s_i of S, we define a modification s' by

$$s'_i = \frac{s_i + \sum_\alpha \varphi_{i\alpha}(S)\pi_{i\alpha}}{1 + \sum_\alpha \varphi_{i\alpha}(S)}$$

and let S' be the n-tuple $S' = (s'_1, s'_2, \ldots, s'_n)$.

We now show that the fixed points of the mapping $\mathbf{T} : S \to S'$ are the equilibrium points.

If S is an equilibrium point, then each $\varphi_{i\alpha}(S) = \max(0, p_{i\alpha}(S) - p_i(S))$ is 0 so that S is a fixed point under $\boldsymbol{T}$.

Conversely, suppose S is fixed under $\boldsymbol{T}$.

For any S (fixed or not), the ith player's mixed strategy s_i will use certain of his pure strategies. Some one of these strategies, say $\pi_{i\alpha}$, must be "least profitable" so that $p_{i\alpha}(S) \le p_i(S)$, which will make $\varphi_{i\alpha}(S) = 0$.

But if S is fixed under $\boldsymbol{T}$, the proportion of $\pi_{i\alpha}$ used in s_i must not be decreased by $\boldsymbol{T}$. Hence, for all β's we must have $\varphi_{i\beta}(S) = 0$ to prevent the denominator of s'_i from exceeding 1.

$$s'_i = \frac{s_i + \sum_\alpha \varphi_{i\beta}(S)\pi_{i\alpha}}{1 + \sum_\alpha \varphi_{i\alpha}(S)}$$

Thus, if S is fixed under $\boldsymbol{T}$, for any i and β, we have $\varphi_{i\beta}(S) = 0$. *But that means that no player can improve his payoff by moving to a pure strategy* $\pi_{i\beta}$. That's exactly the criterion for an equilibrium point.

Hence, equilibrium points correspond to fixed points of a certain continuous function $\boldsymbol{T}$ from the space of all n-tuples of mixed strategies to itself. Note that the function $\boldsymbol{T}$ does not change any of the pure strategies; it only modifies the weights c_{ia} for each player. But the set of all possible weights for each player is a collection of finite dimensional vectors of nonnegative entries that add to 1. In the language of Chapter 9, we are dealing with sets of normalized prices. It should not be a surprise then that domain of $\boldsymbol{T}$ has the fixed-point property. Hence, $\boldsymbol{T}$ must have a fixed point by Brouwer's Theorem, and so Nash's Theorem is proved. ◇

VI. Dynamic Solutions

We conclude our brief introduction to the rapidly expanding discipline of game theory with an examination of how you might play a game where you know the strategies available to you and the other player but you are not told exactly what the payoff matrix is. Suppose, for example, that we have a 2×2 zero-sum game with unspecified payoffs a, b, c, and d as shown in Fig. 16.23.

If Rose and Colin do not know the payoffs, they can only try to guess what an optimal strategy mixture might be. Here is one way to guess. Each tries some initial arbitrary mixed strategy $(x, 1-x)$ for Rose and $(y, 1-y)$ for Colin. When they use this pair of mixed strategies for some time, each gets an average return. These expected returns are

$$\begin{aligned} \text{Rose: EVR} &= axy + bx(1-y) + c(1-x) + d(1-x)(1-y) \\ &= (a-b-c+d)xy + (b-d)x + (c-d)y + d \end{aligned}$$

$$\text{Colin: EVC} = -\text{EVR}$$

Now Rose and Colin only experience seeing the average payoff after a number of plays of the game; they remain ignorant of the values of a, b, c, and d.

Let Rose switch to some other mixed strategy $(x^{\neq}, 1-x^{\neq})$ where $x^{\neq} > x$—that is, Rose chooses R1 more frequently. Her average payoff may increase or decrease. If it increases, then Rose knows she is "on the right track" toward a better mixture, and she will try increase the frequency of R1 even more. If her average payoff decreases, then she will reduce how often she plays R1.

Furthermore, let's assume that if Rose sees a big increase in her average payoff, she will be inclined to make a big change in x. If she experiences a small change in her average payoff, she will make a small change in x. We can model Rose's dynamics as a differential equation

	$C1$	$C2$
$R1$	a	b
$R2$	d	d

FIGURE 16.23 A 2×2 zero-sum game with unspecified numerical payoffs.

FIGURE 16.24 Our first example payoff matrix for the dynamic approach.

	$C1$	$C2$
$R1$	$(5, -5)$	$(-3, 3)$
$R2$	$(-4, 4)$	$(7, -7)$

$$\frac{dx}{dt} = k_1 \frac{\partial EVR(x, y)}{\partial x} \text{ for some } k_1 > 0$$

With similar assumptions about Colin, we obtain

$$\frac{dx}{dt} = k_2 \frac{\partial EVR(x, y)}{\partial y} \text{ for some } k_2 > 0$$

To see how this system of differential equations behaves over time, we examine first the specific 2×2 zero-sum game with payoff matrix as shown in Fig. 16.24.

Here the expected payoff to Rose with mixed strategies $(x, 1-x)$ and $(y, 1-y)$ is

$$5xy - 4y(1-x) - 3x(1-y) + 7(1-x)(1-y) = 19xy - 10x - 11y + 7$$

and since it is zero-sum game, the expected payoff to Colin is

$$-(19xy - 10x - 11y + 7) = -19xy + 10x + 11y - 7$$

Neither Rose nor Colin is aware of these formulas for the expected payoffs. Each just sees the numerical payoffs. If, for example, each played both their strategies exactly half the time, Rose would experience an average payoff of 1.25. If Rose lets $x = .6$ and Colin chooses $y = .3$, then Rose would see an average payoff of 1.12.

[Note that when we analyzed this game earlier in the chapter, we wrote Rose's expected value as

$$19xy - 10x - 11y + 7 = 19\left(x - \frac{11}{19}\right)\left(y - \frac{10}{19}\right) + \frac{23}{19}$$

and concluded that Rose and Colin's optimal choices were $x = 11/19$ and $y = 10/19$.]

The system of differential equations becomes

$$\frac{dx}{dt} = \frac{\partial(19xy - 10x - 11y + 7)}{\partial x} = 19y - 10$$

$$\frac{dy}{dt} = \frac{\partial(-19xy + 10x + 11y + 7)}{\partial y} = -19x + 10$$

The stable point for the system is $x = \frac{11}{19}$, $y = \frac{10}{19}$.

To find the orbit for the system, we find the equation for dy/dx, separate the variables, and integrate. Our successive equations are

$$\frac{dy}{dx} = \frac{dy/dt}{dx/dt} = \frac{-19x + 11}{19y - 10}$$

$$\int (19y - 10)dy = \int (-19x + 11)dx$$

$$\frac{19}{2}y^2 - 10y = -\frac{19}{2}x^2 + 11x + C$$

$$19y^2 - 20y = -19x^2 + 22x + C$$

$$19(y^2 - \frac{20}{19}y) + 19(x^2 - \frac{22}{19}x) = C$$

Now we complete the squares in x and y to find

$$(x - \frac{11}{19})^2 + (y - \frac{10}{19})^2 = \frac{D}{19}$$

for some constant D, depending on the initial values. From this form for the equation of the trajectory, we see that it is a circle with center at

$$\left(\frac{11}{19}, \frac{10}{19}\right)$$

Using a similar approach that we employed in Chapters 2 and 4, we can find explicit solutions of the differential equations. Make the change of variables $X(t) = x(t) - 11/19$ and $Y(t) = y(t) - 10/19$. Then $X' = x' = 19(y - 10/19) = 19Y$ and $Y' = y' = -19(x - 11/19) = -19X$. Hence,

$$X'' = (X')' = (19Y)' = 19Y' = 19(-19X) = -19^2 X$$
$$Y'' = (Y')' = (-19X)' = 19X' = -19(19Y) = -19^2 Y$$

and this system has solutions

$$X(t) = A \sin(19t) + B \cos(19t)$$
$$Y(t) = C \sin(19t) + D \cos(19t)$$

so that

$$x(t) = X(t) + 11/19 = A \sin(19t) + B \cos(19t) + 11/19$$
$$y(t) = Y(t) + 10/19 = C \sin(19t) + D \cos(19t) + 10/19$$

for appropriately chosen constants A, B, C, and D.

Thus, the functions $x(t)$ and $y(t)$ are periodic, but their average values are 11/19 and 10/19, respectively, regardless of what values they initially chose for x and y. Their average

FIGURE 16.25 Our second example payoff matrix for the dynamic approach.

	$C1$	$C2$
$R1$	$(5,-3)$	$(-4,4)$
$R2$	$(-5,5)$	$(3,-4)$

payoffs would be same as using the optimal strategy mixtures. Here Rose and Colin, over the long haul, will do as well in the game as if they knew the entries in the payoff matrix.

In particular, if both players start with a 50-50 mixture of their two strategies $(x(0) = 1/2,\ y(0) = 1/2)$, then

$$x(t) = -\frac{1}{38}\sin(19t) - \frac{3}{38}\cos(19t) + \frac{11}{19}$$

$$y(t) = \frac{1}{38}\sin(19t) - \frac{1}{38}\cos(19t) + \frac{10}{19}$$

As a second example of this technique, we examine the nonzero-sum game whose payoff matrix is displayed in Fig. 16.25:

Here the expected payoff to Rose is $17xy - 7x - 8y + 3$, and the expected payoff to Colin is $-16xy + 8x + 9y - 4$. The differential equations become

$$\frac{dx}{dt} = 17y - 7$$

$$\frac{dy}{dt} = -16x + 9$$

The equation for the trajectory is

$$\left(x - \frac{9}{16}\right)^2 + \left(y - \frac{7}{17}\right)^2 = D$$

from which we see that $(x(t), y(t))$ will travel around a circle whose center is $\left(\frac{9}{16}, \frac{7}{17}\right)$. On average, Rose will play R1 will probability $\frac{9}{16}$ and Colin will play C1 with probability $\frac{7}{17}$. Their long-term expected payoffs will be the same as if they used the mixed strategies $\left(\frac{9}{16}, \frac{7}{16}\right)$ and $\left(\frac{7}{17}, \frac{10}{17}\right)$. Observe now that this game is the same one as displayed in Fig. 16.22; our analysis there showed that these mixed strategies are precisely the Nash equilibrium mixtures. Thus, the dynamic approach yields essentially the same solution as computing the Nash equilibrium when the players know the entries in the payoff matrix.

Finally, let's see what this dynamic approach says about a Prisoners' Dilemma game. Recall the payoff matrix facing Tosca and Scarpia as shown in Fig. 16.26. Tosca's strategies are T1 (Keep the Bargain) and T2 (Double Cross); similarly S1 (Keep the Bargain) and S2 (Double Cross) are Scarpia's strategies. If Tosca uses T1 with probability x and

	$S1$	$S2$
$T1$	$(3,3)$	$(1,4)$
$T2$	$(4,1)$	$(2,2)$

FIGURE 16.26 Our second example payoff matrix for the dynamic approach.

Scarpia employs S1 with probability y, then the expected payoffs are $-x - 2y + 2$ for Tosca and $2x - y + 2$ for Scarpia.

The differential equations of the dynamic approach are

$$\frac{dx}{dt} = \frac{\partial(-x - 2y + 2)}{\partial x} = -1$$

$$\frac{dy}{dt} = \frac{\partial(2x - y + 2)}{\partial y} = -1$$

where both derivatives are negative constants. Thus, no matter what the starting probabilities, both Tosca and Scarpia will use less and less of their first strategy. Over time, x and y will decrease to 0 and the process ends with both using the Double Cross strategy exclusively. The dynamic approach leads to the same outcome as the advice to use a dominant strategy if you have one. Tosca and Scarpia do reach the Nash equilibrium where each receives their second worst outcome. The dynamic approach fails to converge on the available outcome that both Tosca and Scarpia would have favored over the one that happened.

VII. Historical and Biographical Notes

Prior to the publication of von Neumann and Morgenstern's 625-page tome in 1944, there were only a few isolated results in what is now known as game theory. In 1913, Ernst Zermelo (1871–1953) proved that in any finite two-person game of perfect information where players alternate moves and in which chance plays no role, then one of the players must have a *winning strategy*; that is a strategy that insures a victory or a draw. Thus, there is a winning strategy for chess, although no one knows what it is.

In the early 1920s, Emile Borel (1871–1956) published several papers, defining the idea of games of strategy and suggesting that mixed strategies might lead to stable outcomes. Using poker as an example, Borel addressed the tactic of bluffing in games of imperfect information. Von Neumann's 1928 paper stating and proving the Minimax Theorem is generally considered the beginning of modern game theory. In the light of Nash's later work, we can describe von Neumann's result as showing the existence of at least one Nash equilibrium for a two-person zero-sum game and, if there are multiple Nash equilibria, then they all have the same expected payoff.

Von Neumann published nothing more on the subject until Morgenstern began collaborating with him during World War II. In the early 1940s, von Neumann served on many committees and commissions related to the war effort, advised both the army and the navy, and consulted on the ultrasecret Manhattan Project that developed the atomic bomb. How did he have time to coauthor such a large and ambitious book?

Von Neumann's wife Klári recalled how he and Morgenstern worked on the project:

> *Johnny would get home in the evening after having zig-zagged through a number of meetings up and down the coast. As soon as he got in, he called Oskar and then they would spend the better half of the night writing the book. . . . This went on for nearly two years, with continuous interruptions of one kind or the other. Sometimes they could not get together for a couple of weeks, but the moment Johnny got back, he was ready to pick up right where they stopped, as if nothing had happened since the last session. (quoted from Dyson [2012])*

In his own account of the collaboration, Morgenstern [1976] adds more detail and recalls how frustrated von Neumann's wife became over the amount of time the two men spent together:

> *There were endless meetings. . . . We wrote virtually everything together and in the manuscript there are sometimes long passages written by one or the other and also passages in which the handwriting changes two or three times on the same page. We spent most afternoons together, consuming quantities of coffee and Klári was often rather distressed by our perpetual collaboration and incessant conversations . . . she teased us by saying that she would have nothing more to do with the ominous book, which grew larger and larger and consumed more and more of our time[,] if it didn't also have an elephant in it. So we promised we would happily put an elephant in the book.*
>
> *[You can find the elephant if you look closely at page 64 of von Neumann and Morgenstern.]*

The Theory of Games of Economic Behavior drew rave reviews when it first appeared. "Posterity," Arthur Copeland wrote "may regard this book as one of the major scientific achievements of the first half of the twentieth century." From our perspective seven decades later, we see that von Neumann and Morgenstern had not created a fully developed theory that answered all questions, but they did lay a solid foundation. "Nevertheless[,] to the economists and social scientists of the time," Antonia Jones observed, "it must have seen that the answer to their prayers had magically appeared overnight."

The next breakthrough in game theory came in 1950 with John Nash's discovery of equilibrium strategies for n-person noncooperative games. The potential of Nash's idea did not become fully exploited until the early 1970s, when economists discovered how powerful a tool the equilibrium concept could be.

John Nash in his college years.

John Forbes Nash Jr. was born on June 13, 1928, in Bluefield, West Virginia. His father was an electrical engineer and his mother a school teacher. Bluefield, in Nash's words, was "a small city in a comparatively remote geographical location in the Appalachians, not a community of scholars or of high technology." He supplemented his public school education with his own reading, including *Compton's Pictured Encylopedia*, and taking courses at a local community college while in high school.

Nash won a George Washington Scholarship, which paid all his expenses at Carnegie Tech (now Carnegie Mellon University). He began as a chemical engineering student, switched to chemistry, and eventually decided on a mathematics major, in part because the math faculty explained to Nash "that it was not almost impossible to make a good career in America as a mathematician." He was simultaneously awarded both a bachelor's and a master's degree when he graduated at the age of 20. His mathematics professor Richard J. Duffin wrote a one-line letter of recommendation in support of Nash's Princeton graduate school application: "This man is a genius."

The single economics course Nash ever took was an elective class in International Economics, but it had profound consequences. It led directly to the ideas behind his *Econometrica* paper "The Bargaining Problem" and spurred his interest in game theory. As von Neumann and Morgenstern were both in Princeton, the university was a center of activity in this new discipline. Within a year of his arrival on campus, Nash had formulated his idea about equilibrium in noncooperative games and proved the existence of such strategies in n-person games. The impact of his short (27-page) doctoral dissertation was the basis of his 1994 Nobel Prize in Economics.

Nash left Princeton in 1951 to accept a faculty position at the Massachusetts Institute of Technology. Over the next several years, he developed brilliant solutions to several outstanding open research questions in different fields of mathematics. He would hear about major unsolved problems in Riemannian geometry or partial differential equations and then embark on his own unique path to solve them. Before age 30, Nash had earned the "genius" description.

Then tragedy struck. Nash fell victim to paranoid schizophrenia. He had auditory delusions that he believed. He turned down an offer for a professorship at the University of Chicago, for example, claiming that he was shortly to become the emperor of Antarctica. He had many periods of enforced stay at psychiatric hospitals where he was subject to a variety of treatments, including insulin shock therapy—none of which were successful.

There were intermittent periods when Nash was able to overcome the delusions:

> *And it did happen that when I had been long enough hospitalized that I would finally renounce my delusional hypotheses and revert to thinking of myself as a human of more conventional circumstances and return to mathematical research. In these interludes of, as it were, enforced rationality, I did succeed in doing some respectable mathematical research . . .*
>
> *But after my return to the dream-like delusional hypotheses in the later 60s I became a person of delusionally influenced thinking but of relatively moderate behavior and thus tended to avoid hospitalization and the direct attention of psychiatrists.*

Nash eventually returned to Princeton where he was seen as a ghostly, sad figure—"The Phantom of Fine Hall"—who wrote complex equations at night on classroom blackboards. By the early 1970s, Nash began to emerge from schizophrenia, essentially by his own will. "I began to intellectually reject some of the delusionally influenced lines of thinking which had been characteristic of my orientation," he noted in the autobiographical

statement he prepared in conjunction with his receipt of the Nobel award. "So at the present time I seem to be thinking rationally again in the style that is characteristic of scientists. However this is not entirely a matter of joy as if someone returned from physical disability to good physical health. One aspect of this is that rationality of thought imposes a limit on a person's concept of his relation to the cosmos."

Nash's life and his struggle with mental illness formed the basis of a major 2001 motion picture *A Beautiful Mind*, based on Sylvia Nasar's award-winning biography of the same title. Nash associated his madness with living on an "ultralogical" plane, "breathing air too rare" for most mortals, and if being "cured" meant he could no longer do any original work at that level, then, Nash argued, a remission might not be worthwhile in the end. At the very beginning of her biography, Nasar recounts a story that illustrates the often fine line between genius and madness, the belief that Nash apparently had that original creative ideas may come from the same part of the mind that generates delusions. It is May 1959, and Harvard mathematician George Mackey is visiting Nash in the psychiatric hospital:

> *Nash was slumped in an armchair in one corner of the hospital lounge, carelessly dressed in a nylon shirt that hung limply over his unbelted trousers. His powerful frame was slack as a rag doll's, his finely molded features expressionless. He had been staring dully at a spot immediately in front of the left foot of Harvard professor George Mackey, hardly moving except to brush his long dark hair away from his forehead in a fitful, repetitive motion. His visitor sat upright, oppressed by the silence, acutely conscious that the doors to the room were locked. Mackey finally could contain himself no longer. His voice was slightly querulous, but he strained to be gentle. "How could you," began Mackey, "how could you, a mathematician, a man devoted to reason and logical proof . . . how could you believe that extraterrestrials are sending you messages? How could you believe that you are being recruited by aliens from outer space to save the world? How could you . . . ?"*
>
> *Nash looked up at last and fixed Mackey with an unblinking stare as cool and dispassionate as that of any bird or snake. "Because," Nash said slowly in his soft, reasonable southern drawl, as if talking to himself, "the ideas I had about supernatural beings came to me the same way that my mathematical ideas did. So I took them seriously."*

The Nash equilibrium, writes Robert Aumann, also a Nobel Prize winner in economics for his work in game theory, "is without doubt the single game theoretic solution concept that is most frequently applied in economics. Economic applications include oligopoly, entry and exit, market equilibrium, search, location, bargaining, product quality, auctions, insurance, principal-agent [problems], higher education, discrimination, public goods, what have you. On the political front, applications include voting, arms control and inspection, as well as most international political models (deterrence, etc.). Biological applications all deal with forms of strategic equilibrium; they suggest an interpretation of equilibrium quite different from the usual overt rationalism."

EXERCISES

1. Suppose David thought there was a probability of .8 that he could beat Goliath and that he assigned utility values to L and W as -10 and $+90$. How large would M have to be for David to decide not to fight?

2. A physician advises a patient suffering from angina that his best options for treatment is bypass surgery, which has a 85% of being successful and relieving him of pain; unfortunately, there is 15% chance that the patient

will die during the operation. Without the operation, the patient can expect to live for many years but with recurring chest pain. Suppose the patient's utility scale ranges from 0 (worst outcome) to 1 (best outcome). Should he elect the surgery if he evaluates continuing to live with angina as having a utility of .7?

3. The mayor of the largest city in her state has just won her party's nomination to run for governor. The mayor needs to decide whether she should resign her position and campaign full-time for governor or stay on the job and campaign part-time. She estimates that she has a .65 probability of being elected if she devotes full time to campaigning, but only a .55 chance with a part-time effort. The worst outcome (utility 0) would be to resign as mayor and lose the governor's race; she would then have to find a new job. The best outcome (utility 1) is becoming governor. Let m be utility of remaining mayor after a part-time unsuccessful governor's campaign. How small a value of m would it take for the mayor to decide to quite her current job and devote all her time trying for the governorship?

4. Create a payoff matrix for the *Battle of the Sexes* game.

5. In our example of the wealthy man with three daughters, we created a scenario in which the older and younger daughter are each vying to convince the middle daughter to split the estate two ways.

(a) The younger daughter has just offered the middle one a (0, .6, .4) split. If you are the older daughter, what would be your counteroffer?

(b) If these negotiations continue, the middle daughter will be offered a sequence of offers of the form $(0, m, 1-m)$ or $(1-m, m, 0)$ with an ever-increasing m. How high do you think m might get before the oldest daughter offers the youngest a $(.5, 0, .5)$ division of the money? What might happen next?

6. Write out in words the optimal strategy for playing tic-tac-toe.

7. Suppose the payoffs in Rock-Paper-Scissors are $+1$ for a winner, -1 for the loser, and 0 to each if it is a tie. Write out the payoff matrix. What is the lower value? The upper value? Is there a saddle point?

8. In the game Matching Pennies, Rose and Colin each has a penny, which he or she simultaneously places on the table. If the pennies match (both heads or both tails), Rose keeps both pennies, but if the pennies do not match (one head and one tail), Colin keeps both.

(a) Write out the payoff matrix for this 2×2 zero-sum game.

(b) What is the optimal way to play this game?

(c) What is the expected payoff of the game if both Rose and Colin play optimally?

9. A zero-sum game is *fair* if the expected payoff of optimal strategies is 0. Is Matching Pennies fair? Is Rock-Paper-Scissors fair? Give an example of a game that is not fair.

10. Find all dominating rows and columns for the zero-sum game with matrix

	$C1$	$C2$	$C3$	$C4$
$R1$	9	3	12	3
$R2$	3	0	-6	-63
$R3$	6	3	9	3
$R4$	-51	-3	45	0

11. Find lower value, upper value, and saddle points (if any) for the game shown in Exercise 10.

12. Consider the zero-sum game with payoff matrix

	$C1$	$C2$	$C3$	$C4$
$R1$	17	27	7	-3
$R2$	27	47	17	37
$R3$	-3	7	27	17
$R4$	-13	-3	-3	7

(a) Show that R2 dominates R1 and also dominates R4 so that we can reduce the game by deleting rows 1 and 4.

(b) In the resulting 2×4 game, show that C1 dominates C2 and C4 so that columns 2 and 4 can be deleted.

(c) Show that the resulting 2×2 game has the payoff matrix

	$C1$	$C3$
$R2$	4	2
$R3$	-2	4

(d) Find the lower and upper values for the 2×2 game. Is there a saddle point?

(e) Determine the optimal mixed strategies for the game.

13. Examine the payoff matrix for a zero-sum game

	$C1$	$C2$	$C3$	$C4$
$R1$	7	5	9	5
$R2$	5	3	9	1
$R3$	5	5	5	5

(a) Show that this game has a value 5.

(b) Show that there are four saddle points for this game.

(c) Show that not every entry of 5 in the matrix is a saddle point.

(d) Show that a saddle point occurs in a dominated strategy.

14. [Straffin] The payoff matrix in Exercise 13 raises a concern. Since a saddle point can appear in a dominated row and we have advised never using a dominated row, it is conceivable that eliminating a dominated row might remove the only saddle point in a game. Show that this actually cannot happen: prove that the Dominance Principle and the Saddle Point Principle cannot conflict with each other. [Hint: Prove that if row A dominates row B and row B contains a saddle point of the game, then the entry in row A in the same column of row B that holds a saddle point is itself a saddle point.]

15. For the zero-sum game with payoff matrix

	$C1$	$C2$
$R1$	1	3
$R2$	4	2

, find

(a) Lower value and upper value

(b) Optimal mixed strategies for Rose and Colin

(c) The value of the game

16. Repeat Exercise 15 if the payoff matrix is

	$C1$	$C2$
$R1$	-5	4
$R2$	6	0

17. For the zero-sum game with payoff matrix

	$C1$	$C2$
$R1$	a	b
$R2$	c	d

where $D = a - b - c + d \neq 0$, show that the optimal strategy for Rose is to play R1 with probability $(d - c)/D$ and for Colin is to play C1 with probability $(d - b)/D$. What should the players do if $D = 0$?

18. In our game theory model of Abraham and the Sacrifice of Isaac, we assumed that God's preference ordering was

$$OR > OR^* > O^*R^* > O^*R$$

Determine the payoff matrix and the optimal strategy choices for both players—for each of the three different orderings for Abraham—if God's preferences are given by

(a) $OR > O^*R > OR^* > O^*R^*$

(b) $OR > OR^* > O^*R > O^*R^*$

19. Investigate another plausible ordering for God's preferences and analyze the payoff matrices for each of the three given orderings of Abraham's preferences.

20. Describe and justify what you consider to be the most realistic preference orderings for God and Abraham, then determine the payoff matrix and find the optimal strategy choices for the two players.

21. Is there any example of preference orderings you consider to be reasonable ones where Abraham's rational strategy is anything other than to offer Isaac as a sacrifice? Do you believe that God designed a good test for Abraham's faith? Explain.

22. Show that in our model of the biblical story Abraham has 24 different possible rankings of the four outcomes OR, O^*R, OR^*, and O^*R^*.

23. List all 24 possible rankings for Abraham.

24. For each of the 24 rankings for Abraham, determine if he has a dominant strategy.

25. In analyzing the "sacrifice" story of Jephthah (Judges 11), Steven Brams assumes that Jephthah and God have the same strategies available as Abraham and God had in the Genesis episode. He considers the same three rankings for Jephthah as we earlier considered for Abraham: faithful, somewhat wavering, seriously wavering. But he considers two different rankings for God:

$OR^* > OR > O^*R^* > O^*R$ (show-of-faith interpretation of God's preferences)
$OR^* > O^*R^* > OR > O^*R$ (vindictive interpretation of God's preferences)

For each of these two rankings, determine the payoff matrix, find any dominant strategies, and predict the outcome of the game. How does your prediction compare with what happened in the Bible?

26. **(a)** For the game in Fig. 16.25, show that the expected payoff for Rose is $\mathrm{EVR}(p,q) = 17pq - 7p - 8q + 3$ and the expected payoff for Colin is $\mathrm{EVC}(p,q) = -16pq + 8p + 9q - 4$ if Rose chooses R1 with probability p and Colin chooses C1 with probability q.

(b) Demonstrate that EVR and EVC have the properties that $\mathrm{EVR}(p, 7/17) = -1/5$ for all p and $\mathrm{EVC}(9/16, q) = ½$ for all q.

27. **(a)** Show that both R1C1 and R2C2 are Nash equilibrium for the nonzero-sum game

	$C1$	$C2$
$R1$	(3, 2)	(1, 1)
$R2$	(2, 2)	(2, 3)

(b) Is either R1C2 or R2C1 a Nash equilibrium?

28. Find the mixed strategy Nash equilibrium if the payoff matrix is

	$C1$	$C2$
$R1$	(4, 8)	(2, 0)
$R2$	(6, 0)	(0, 8)

29. Find the mixed strategy Nash equilibrium if the payoff matrix is

	$C1$	$C2$
$R1$	(4, 7)	(1, 0)
$R2$	(6, 2)	(0, 8)

30. Is there a mixed strategy Nash equilibrium for

(a) The game of chicken?

(b) The game Tosca and Scarpia play?

31. Determine whether there are Nash equilibria in pure strategy choices for the game with payoff matrix

	$C1$	$C2$	$C3$	$C4$
$R1$	(9, 1)	(9, 1)	(0, 0)	(0, 0)
$R2$	(5, 3)	(4, 4)	(5, 3)	(4, 4)

32. Prove that $\max_{\text{all } r'_i\text{s}} [p_i(S; r_i)] = \max_{\alpha} [p_i(S; \pi_{i\alpha})]$.

33. Give a complete justification for the claim that a necessary and sufficient condition for S to be an equilibrium point is

$$p_i(S) = \max_{\text{all } r'_i\text{s}} [p_i(S; r_i)] = \max_{\alpha} [p_i(S; \pi_{i\alpha})] = \max_{\alpha} p_{i\alpha}(S)$$

34. Explain why the transform **T** is continuous.

35. Modify, if necessary, the proof of Nash's theorem to give a proof of von Neumann's Minimax Theorem.

36. Use the dynamic approach to analyze the game of Exercise 15.

37. Use the dynamic approach to analyze the game of Exercise 16.

38. What outcome does the dynamic approach predict for the game with payoff matrix

	$C1$	$C2$
$R1$	1	4
$R2$	5	8

?

39. What outcome does the dynamic approach predict for the game of Exercise 27?

40. Here is a game with infinitely many strategies: Rose and Colin each picks one positive integer. If both pick the same integer N, then Colin pays $f(N)$ dollars to Rose for some given payoff function f. If they choose different integers, then no money is exchanged; both get \$0.

(a) Why does each player have infinitely many strategies?

(b) How should they play the game if $f(N) = (N-3)^2$?

(c) If the function f takes on both positive and negative values, show that each player has a strategy that results in both getting a payoff of 0 and that the value of the game is 0.

(d) Show that in an equilibrium, each of Rose's pure responses to a mixed strategy by Colin yields the same payoff.

(e) If every value of $f(N)$ is positive, so that an equilibrium strategy has a player choose positive integer N with probability $\frac{1}{f(N)}\left(\sum_{k=1}^{\infty}\frac{1}{f(k)}\right)^{-1}$ if the infinite sum converges.

(f) Suppose $f(N) = N^2$ for all N. Use the fact that $\sum_{k=1}^{\infty}\frac{1}{k^2} = \frac{\pi^2}{6}$ to show that the equilibrium strategy chooses N with probability $\frac{1}{N^2}\frac{6}{\pi^2}$ and that the value of this game is $\frac{6}{\pi^2}$.

SUGGESTED PROJECTS

1. In some games against nature, we may have no knowledge of the probabilities with which each state of nature may occur. How do you play such a one-person game? Assume all of nature's states are equally likely? Use a maximin approach? Are there other ideas with interesting properties? How much should you be willing to pay to get good estimates of nature's probabilities? A good place to begin an investigation of such games is a classic 1951 paper by John Milnor, "Games against Nature," reprinted in Martin Shubik, ed., *Game Theory and Related Approaches to Social Behavior*, New York: Wiley, 1964.

2. Decision theory is now a well-developed discipline analyzing one-person games in which the states of nature have known probabilities. Behn and Vaupel [1982] provide an easy introduction to this intriguing topic, which has many important applications you can analyze.

3. Duels provide another source with infinitely many strategies. Two duelists start a fixed distance apart and walk toward each other. Each must decide at what distance between them they will fire. More generally, a duel is a two-person game in which each player can take as long as she likes to prepare his or her move, but the other player can take advantage of his or her hesitation. Dresher [1961] is a good place to start. See also D. Marc Kilgour and Steven J. Brams, "The Truel," *Mathematics Magazine* 70 (1997): 315–326, for a study of duels among three opponents.

4. The Nash equilibrium is but one idea for what is the appropriate notion of a "solution" for a general *n*-person game. Other concepts worth exploring are stable set, core, nucleolus, and kernel imputation among others. Anatol Rapoport, n-*Person Game Theory: Concepts and Applications*, New York: Dover, 2013, provides an enlightening introduction.

5. As we have seen, the Nash equilibrium solution does not always provide the most desirable payoffs to the players. There are also difficulties when multiple Nash equilibria exist. Investigate Robert Aumann's idea of a *correlated equilibrium* to deal with such problems. See Aumann [1974, 1987].

6. Explore other applications of game theory to religious texts. Aumann and Maschler [1985], for example, show evidence that the nucleolus appears in an ancient Talmudic text. See also O'Neill [1982] and Brams [1980, 2007].

7. Instead of playing a game only once, we may have the opportunity to play against the same person repeatedly. So-called *iterative* games enable each player to "signal" to the other a willingness to cooperate or to retaliate in later plays of the game by how they play in early rounds. The "best" strategy in a repeated game may be quite different from what it is for a single play. See Axelrod [2006], Aumann and Maschler [1995], and Sorin [2002] for more on the theory and applications.

8. Another major result in game theory due to John Nash [1950] is his solution to "The Bargaining Problem," a two-person cooperative game where the players can make binding agreements on how to play. Nash's paper derives a solution from a plausible set of axioms.

9. A *characteristic function* for an *n*-person game is a function that assigned a numerical value to each coalition (subset) of the players. Lloyd Shapley (1923–), another Nobel Prize winner, derived a solution concept for such games from a short list of appealing axioms. Investigate Shapley's approach [1953] and some of the many applications of what is now called the *Shapley Value*. Sergiu Hart maintains a bibliography on value theory at http://www.ma.huji.ac.il/hart/value.html

10. How much relative power do the president, the Senate (100 members) and the House of Representatives (435 members) have in enacting legislation? One way a bill may become a law is with the approval of a majority of both houses and the president's approval. Another path is by the approval of at least two-thirds of both houses over the objection of the president. Shapley and Martin Shubik (1926–) developed an index of power in voting games that showed that power is not always proportional to size; it has more to do with how often coalitions cast deciding votes. See Shapley and Shubik's 1954 paper.

You can find a listing of references and suggestions for additional reading on the books's website, www.wiley.com/college/olinick

Sets

APPENDIX I

By a *set* we mean a well-defined collection of objects, called the *elements* or (members) of the set. Some examples of sets are

1. The set A of real numbers less than 21

2. The set B of college sophomores in Texas universities

3. The set C of negative integers

4. The set D of three-headed residents of Muskegon, Michigan

5. The set E of solutions of the equation $\tan x - \log x = x^3$

6. The set F of integers strictly between 3 and 10

We use the notation "$x \in X$" to represent the statement that "The element x is a member of the set X." If x is not an element of X, denote this by $x \notin X$. In our examples, $4 \in A$ and $24 \notin A$.

Sets may be described in terms of some common property shared by the elements. A set may also be given by listing all its members; when this is done, the elements are typically written within braces. Here are some further examples:

7. $G = \{\text{single, double, triple, home run}\}$

8. $H = \{4, 5, 6, 7, 8, 9\}$

9. $I = \{1, 2, 3, \ldots\}$

10. $J = \{\text{Washington, Adams, Jefferson, } \ldots, \text{ Clinton, Bush, Obama}\}$

Some sets occur so frequently in applications that special symbols have been invented for them. The set of real numbers, for example, is commonly denoted by $\mathbb{R}$ and the set of integers by $\mathbb{Z}$.

A third way of describing a set is by a special notation easily understood by an example. The set of integers strictly between 3 and 10 would be written

$$\{x \in \mathbb{Z} : 3 < x < 10\}$$

where the colon ":" is read "such that." Note that the descriptions of the sets F and H in these examples specify the same collection of numbers. So does the description $K = \{9, 8, 7, 6, 5, 4\}$. It is reasonable to call these sets *equal* by the following definitions.

DEFINITION If X and Y are sets, then X and Y are *equal*, denoted $X = Y$, precisely if the sets contain exactly the same elements.

The sets F and A of the examples are not equal, since 17 is an element of A but not of F. However, every element of F is an element of A, so F is a subcollection, or *subset*, of A.

DEFINITION If X and Y are sets and every element of X is also an element of Y, then we say that X is a *subset* of Y. This relationship is denoted by $X \subseteq Y$.

Observe that although sets A and I have some common elements, neither is a subset of the other. We might denote this by $A \not\subset I$ and $I \not\subset A$.

***Proposition* 1** $X = Y$ if and only if $X \subseteq Y$ and $Y \subseteq X$.

The set containing no elements is called the *empty set* and is denoted $\varnothing$. Observe that the set D in the examples is the empty set. If X is any set, then $\varnothing \subseteq X$ (proof: try to find some element of $\varnothing$ that is not an element of X). Also note that every set is a subset of itself.

If X is any set, we may consider the collection of all subsets of X. This set is called the *power set* of X. As an example, the set $X = \{x, y, z\}$ has eight distinct subsets:

$$\varnothing, \{x\}, \{y\}, \{z\}, \{x, y\}, \{x, z\}, \{y, z\}, X.$$

A set X is *finite* if it contains exactly n distinct elements for some nonnegative integer n. The sets B, D, E, F, G, H, and J of the examples are finite sets.

***Proposition* 2** If X is a finite set with precisely n distinct elements, then the power set of X contains 2^n distinct elements.

DEFINITION If S is a set and X is a subset of S, then the set of elements of S that are not in X is called the *complement of X in S*. It is denoted $S - X$ or sometimes by X^C (if there is no ambiguity about S).

As an example, if $S = \{1, 2, 3, 4, 5\}$ and $X = \{3, 4\}$, then $S - X = \{1, 2, 5\}$.

Exercise Show that $S - (S - X) = X$.

By definition, the sets X and $S - X$ have no elements in common. This is also true of the sets A and B of the examples and for many other pairs of sets. Such sets are said to be *pairwise disjoint*. Other pairs of sets do share common elements, and there is a special notation for this set of common elements.

DEFINITION If X and Y are sets, then the set $X \cap Y$ is the set of all elements that are in both X and Y. The set $X \cap Y$ is called the *intersection* of X and Y.

In the examples,

$$\begin{aligned} A \cap I &= \{x : x \in A \text{ and } x \in I\} \\ &= \{x : x \text{ is a real number less than } 21 \text{ and } x \text{ is a positive integer}\} \\ &= \{x : x \text{ is a positive integer less than } 21\} \\ &= \{1, 2, \ldots, 20\}. \end{aligned}$$

If X and Y are pairwise disjoint, then we have $X \cap Y = \varnothing$.

There is another important operation of combining sets that consists in forming the collection of all elements that belong to either set.

DEFINITION If S is a set and X and Y are subsets of S, then the *union* of X and Y, denoted $X \cup Y$, is the set of elements belonging to X or Y or both. In our notation,

$$X \cup Y = \{z : z \in X \text{ or } z \in Y\}$$

As an example, suppose that $X = \{1, 2, 3, 5\}$ and $Y = \{3, 4, 5, 6, 7\}$. Then the union is $X \cup Y = \{1, 2, 3, 4, 5, 6, 7\}$.

To study relations and functions, it is necessary to introduce the concept of a Cartesian product of two sets.

DEFINITION If X and Y are any two sets, then the *Cartesian product* of X and Y, denoted $X \times Y$, is the set of all ordered pairs (x, y) where x is a member of X and y is a member of Y. In terms of our notation,

$$X \times Y = \{(x, y) : x \in X \text{ and } y \in Y\}.$$

The next example should clarify this definition. If $X = \{1, 2, 3\}$ and $Y = \{3, 4\}$, then we have

$$X \times Y = \{(1, 3), (1, 4), (2, 3), (2, 4), (3, 3), (3, 4)\}$$
$$Y \times X = \{(3, 1), (3, 2), (3, 3), (4, 1), (4, 2), (4, 3)\}$$

$$Y \times Y = \{(3, 3), (3, 4), (4, 3), (4, 4)\}$$

What is $X \times X$?

The principle of mathematical induction, which is used in a number of proofs in this text, can be formulated in terms of sets. Let $\boldsymbol{N}$ be the set of positive integers, and suppose that X is a subset of N. The *Axiom of Mathematical Induction* asserts,

IF: (i) $1 \in X$ and (ii) whenever $n \in X$, then $n + 1 \in X$

THEN: $X = N$

An equivalent axiom, easier to remember but sometimes more cumbersome to use, is as follows: Every nonempty set of positive integers has a least element.

Example

Show that $1 + 2 + 3 + \cdots + k = \frac{k(k+1)}{2}$ for every positive integer k.

Solution I:

Let X be the subset of $\boldsymbol{N}$ for which the equation is valid. Then $1 \in X$, since $1 = \frac{1(1+1)}{2}$; thus, (i) is satisfied. To show that (ii) is true, suppose that $n \in X$. Then $1 + 2 + \cdots + n = \frac{n(n+1)}{2}$. Adding $n + 1$ to each side of this equation produces

$$\begin{aligned} 1 + 2 + \cdots + n + (n+1) &= (1 + 2 + \cdots + n) + (n+1) \\ &= \frac{n(n+1)}{2} + (n+1) \\ &= \frac{(n+1)(n+2)}{2} \end{aligned}$$

which is just the statement that $n + 1 \in X$. Since (i) and (ii) are true, the principle of mathematical induction asserts that $X = \boldsymbol{N}$—that is, the equation is true for every positive integer.

Solution II:

Let A be the set of all positive integers for which the equation is not true. If A is nonempty, then it has a smallest element k. But then we have

a. $k \neq 1$
b. $1 + 2 + \cdots + k \neq \frac{k(k+1)}{2}$
c. $1 + 2 + \cdots + (k-1) = \frac{(k-1)k}{2}$

Condition (c) is true because $k - 1$ is smaller than k and the equation is true for all positive integers less than k. Now (a), (b), and (c) are inconsistent. We have a contradiction to the assumption that A is nonempty. Thus, A must be empty, and the equation is valid for every positive integer.

EXERCISES

1. Use mathematical induction to prove Proposition 2.

2. Show that $X \cup Y = Y \cup X$ and $X \cap Y = Y \cap X$ for all sets X and Y, but that, in general, $X \times Y \neq Y \times X$.

3. Show that $(X \cup Y) \cup Z = X \cup (Y \cup Z)$ and $(X \cap Y) \cap Z = X \cap (Y \cap Z)$ for all sets X, Y, Z so that the expressions $X \cup Y \cup Z$ and $X \cap Y \cap Z$ are well defined. What can you say about $(X \times Y) \times Z$ and $X \times (Y \times Z)$?

Matrices

APPENDIX II

By a **matrix**, we simply mean a rectangular array of numbers. Examples of matrices are

$$A = \begin{pmatrix} 3 & 2 & 1 \\ -1 & 0 & 7 \end{pmatrix}$$

$$B = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$$

$$C = \begin{pmatrix} .2 & .3 & .5 & .4 \end{pmatrix}$$

$$D = \begin{pmatrix} 1.5 \\ -6 \\ 9.9 \\ 0 \\ 22 \end{pmatrix}$$

Matrices are classified according to their size and shape by specifying the number of their rows and columns. An $m \times n$ *matrix* is a matrix with m rows and n columns. Thus, A is a 2×3 matrix, B is a 2×2 matrix, C is a 1×4 matrix, and D is a 5×1 matrix. Any $1 \times n$ matrix is called a *row vector*, while an $m \times 1$ matrix is said to be a *column vector*. The individual numbers in vectors are called *components* of the vector. If $m = n$, the matrix is said to be *square* and to have an *order* equal to the number of rows. Thus, B is a square matrix of order 2. Two matrices have the same *size* if they have the same number of rows and columns. The four matrices A, B, C, and D, it should be noted, are of different sizes.

The general $m \times n$ matrix M has the form

$$M = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & & & \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix}$$

The number in the ith row and jth column of a matrix is called the ijth entry of the matrix, and may be denoted M_{ij}. Thus, $A_{23} = 7$.

Two matrices are said to be *equal* matrices if they are of the same size and the corresponding entries are all equal. We write $A = B$ to denote that two matrices A and B are equal. For example, if we consider the six matrices:

$$U = (1, 2) \quad V = \begin{pmatrix} 1 \\ 2 \end{pmatrix} \quad W = (1, 2) \quad X = (2, 1) \quad Y = (1, 0) \quad Z = (1, 2, 0)$$

only the matrices U and W are equal.

Matrix addition is defined for matrices of the same size by the addition of the corresponding entries. Thus, the ijth entry of the sum of two matrices is the sum of their ijth entries; in symbols,

$$(A + B)_{ij} = A_{ij} + B_{ij}$$

As an example, if the matrices A and B are given by

$$A = \begin{pmatrix} 3 & 2 & 1 \\ -1 & 0 & 7 \end{pmatrix} \quad \text{and} \quad B = \begin{pmatrix} 5 & 4 & -6 \\ 2 & 1 & 0 \end{pmatrix}$$

then the sum, $A + B$, is given by

$$A + B = \begin{pmatrix} 3+5 & 2+4 & 1-6 \\ -1+2 & 0+1 & 7+0 \end{pmatrix} = \begin{pmatrix} 8 & 6 & -5 \\ 1 & 1 & 7 \end{pmatrix}$$

The matrices A and C, given by

$$A = \begin{pmatrix} 3 & 2 & 1 \\ -1 & 0 & 7 \end{pmatrix} \quad \text{and} \quad C = \begin{pmatrix} 4 & 5 \\ 2 & 0 \\ -1 & -1 \end{pmatrix}$$

cannot be added, even though they have the same number of entries.

A matrix, all of whose entries are 0, is called a *zero matrix*, and we denote the $m \times n$ zero matrix by 0^{mn} or sometimes simply by $\mathbf{0}$ if the size of the matrix is clear from the context.

If A is a matrix and c is a constant, then we can define the *scalar multiple* cA to be the matrix obtained by multiplying each entry of A by the constant c. In symbols, this is $(cA)_{ij} = c(A_{ij})$; that is, the ijth entry of cA is c times the ijth entry of A.

If A is the matrix given above then the matrices $2A$ and $-1A$ are given by

$$2A = \begin{pmatrix} 6 & 4 & 2 \\ -2 & 0 & 14 \end{pmatrix} \quad \text{and} \quad -1A = \begin{pmatrix} -3 & -2 & -1 \\ 1 & 0 & -7 \end{pmatrix}$$

We will denote the matrix $-1A$ simply by $-A$. If A and B are matrices of the same size, then the expression $B + -A$ is well defined as a matrix addition. We will write such an expression as $B - A$ and call the operation *matrix subtraction*. What this idea boils down to is that subtraction of matrices is defined by subtraction of corresponding entries.

Our first theorem lists the basic properties of matrix addition and scalar multiplication. These properties follow easily from analogous properties of the addition and multiplication of ordinary numbers.

THEOREM 1 Let A, B, and C be any $m \times n$ matrices, and let c and d be any real numbers. Then

1. $A + B$ is an $m \times n$ matrix
2. $A + B = B + A$
3. $A + (B + \mathrm{C}) = (A + B) + \mathrm{C}$
4. $A + \mathbf{0} = A$
5. For each matrix A, there is a matrix, $-A$, such that $A + (-A) = \mathbf{0}$
6. cA is an $m \times n$ matrix
7. $c(A + B) = cA + cB$
8. $(c + d)A = cA + dA$
9. $c(dA) = (cd)A$
10. $1A = A$

Proof of Theorem 1 We will prove (2); the other properties follow by similar reasoning and are left as exercises for the reader. By definition of matrix addition, the ijth entry of $A + B$ is $A_{ij} + B_{ij}$. But A_{ij} and B_{ij} are numbers, so we have $A_{ij} + B_{ij} = B_{ij} + A_{ij}$. Now $B_{ij} + A_{ij}$ is the ijth entry of $B + A$. Since the corresponding entries of $A + B$ and $B + A$ are equal, the matrices are equal. ◇

Knowledge of Properties 1–10 of Theorem 1 is essential for working with matrices. You will have little difficulty remembering them as they are so similar to the operations involving real numbers. Matrix multiplication, which we introduce next, is a different story.

Matrix Multiplication

One might define the product of two matrices to be the matrix obtained by multiplying corresponding entries. Such a definition would have a number of applications; you will be invited to explore the consequences of such a definition in the exercises. When mathematicians speak of matrix multiplication, however, they have a different operation in mind, an operation that was invented to handle many very useful applied problems.

To explain this operation, we will begin with the example of a cashier at the checkout counter of the campus bookstore. Suppose that you purchase 6 pencils, 4 notebooks, 2 packs of index cards, 36 paper clips, and 1 sweatshirt at the store. How does the cashier determine what to charge you?

The cashier makes the following calculation:

$$\begin{aligned}\text{Total Cost} &= (\text{Total cost of pencils}) + (\text{total cost of notebooks}) \\ &\quad + (\text{total cost of index cards}) + (\text{total cost of paper clips}) \\ &\quad + (\text{total cost of sweatshirts}) \\ &= (\text{Number of pencils})(\text{cost per pencil}) \\ &\quad + (\text{number of notebooks})(\text{cost per notebook}) \\ &\quad + (\text{number of packs of index cards})(\text{cost per pack}) \\ &\quad + (\text{number of paper clips})(\text{cost per clip}) \\ &\quad + (\text{number of sweatshirts})(\text{cost per shirt})\end{aligned}$$

Let us represent the purchases by a row vector:

$$\begin{aligned}A &= (6\text{ pencils},\ 4\text{ notebooks},\ 2\text{ packs of cards},\ 36\text{ clips},\ 1\text{ sweatshirt}) \\ &= (6,\ 4,\ 2,\ 36,\ 1)\end{aligned}$$

and represent the unit cost of each item by a column vector:

$$B = \begin{pmatrix} 5 \\ 75 \\ 30 \\ 2 \\ 398 \end{pmatrix} \begin{matrix} \text{cents per pencil} \\ \text{cents per notebook} \\ \text{cents per pack of index cards} \\ \text{cents per paper clip} \\ \text{cents per sweatshirt} \end{matrix}$$

Then the total cost is given by

$$\begin{aligned}\text{Total cost} &= (6)(5) + (4)(75) + (2)(30) + 36(2) + 1(398) \\ &= 30 + 300 + 60 + 72 + 398 \\ &= 860 \text{ cents or } \$8.60\end{aligned}$$

Matrix multiplication will be defined so that the total cost is the product of the purchase vector and the cost vector.

As a second example, consider a gamble with three possible outcomes, $+\$5$, $-\$2$, and $+\$25$, with respective probabilities of .25, .7, and .05. Then the expected value of the gamble (see Chapter 10) is $(5)(.25) + (-2)(.7) + (25)(.05) = \$1.10$. If we denote the outcomes of the gamble by the row vector $\mathbf{A} = (5,\ -2,\ 25)$ and the probabilities by the column vector

$$B = \begin{pmatrix} .25 \\ .7 \\ .05 \end{pmatrix}$$

then the expected value is given by the product of these two matrices.

With these two examples in mind, we are ready to make our first formal definition about matrix multiplication.

DEFINITION If A is a $1 \times m$ row vector and B is a $m \times 1$ column vector, then the *product AB* is defined to be the number given by

$$AB = \sum_{k=1}^{m} a_{1k} b_{k1} = a_{11} b_{11} + a_{12} b_{21} + \cdots + a_{1m} b_{m1}$$

where $A = (a_{11}, a_{12}, \ldots, a_{1m})$ and

$$B = \begin{pmatrix} b_{11} \\ b_{21} \\ \cdots \\ b_{m1} \end{pmatrix}$$

Note that a row vector and a column vector may be multiplied by this definition only if the number of components in each vector is the same.

Now suppose that **A** is a $1 \times m$ vector and B is an $m \times n$ matrix. According to the definition given above, it is possible to multiply the row vector **A** by each column of the matrix B simply by treating that column as a column vector. This gives us a natural way to define the product of a vector and a matrix.

DEFINITION If A is a $1 \times m$ row vector and B is a $m \times n$ matrix, then the product AB is defined to be the $1 \times n$ matrix whose jth component is the product of the vector A and the jth column of B.

EXAMPLE

Suppose $A = (3, 2, -1)$ and $B = \begin{pmatrix} 5 & 4 \\ -2 & 0 \\ 1 & 9 \end{pmatrix}$

Then the product AB is a 1×2 matrix. The first component is the product of the vector $(3, 2, -1)$ and the vector

$$\begin{pmatrix} 5 \\ -2 \\ 1 \end{pmatrix}$$

which is $(3)(5) + (2)(-2) + (-1)(1) = 10$. The second component is the product of A and the second column of B. The value of this second component is $(3)(4) + (2)(0) + (-1)(9) = 3$. Thus, the product AB is the matrix $(10, 3)$.

Finally, suppose that A is a $k \times m$ matrix and B is an $m \times n$ matrix. Then it is possible to multiply each row of A by the matrix B by using the definition we have just given. Each such multiplication yields a $1 \times n$ row vector. Fitting these k row vectors together in a natural fashion gives a $k \times n$ matrix.

DEFINITION If A is a $k \times m$ matrix and B is an $m \times n$ matrix, then the *product* AB is defined to be the $k \times n$ matrix whose ijth entry is the product of the ith row of A and the jth column of B. Thus, $(AB)_{ij} = \sum_{r=1}^{m} a_{ir} b_{rj}$.

EXAMPLE

Consider the following four matrices:

$$A = \begin{pmatrix} 1 & 2 & 7 \\ -3 & 0 & 8 \end{pmatrix} \quad B = \begin{pmatrix} 9 & -6 \\ -2 & 1 \\ 4 & 1 \end{pmatrix} \quad C = \begin{pmatrix} 3 & 6 \\ 2 & 4 \end{pmatrix} \quad D = \begin{pmatrix} 2 & 8 \\ 6 & 9 \end{pmatrix}$$

Then we have the following products:

$$AB = \begin{pmatrix} 33 & 3 \\ 5 & 26 \end{pmatrix} \qquad BA = \begin{pmatrix} 27 & 18 & 15 \\ -5 & -4 & -6 \\ 1 & 8 & 36 \end{pmatrix}$$

$$BD = \begin{pmatrix} -18 & 18 \\ 2 & -7 \\ 14 & 41 \end{pmatrix} \qquad DB \text{ is not defined}$$

$$CD = \begin{pmatrix} 42 & 78 \\ 28 & 52 \end{pmatrix} \qquad DC = \begin{pmatrix} 22 & 44 \\ 36 & 72 \end{pmatrix}$$

$$CC = \begin{pmatrix} 21 & 42 \\ 14 & 28 \end{pmatrix} \qquad C(CC) = (CC)C = \begin{pmatrix} 146 & 294 \\ 98 & 196 \end{pmatrix}$$

Note that the product BD is defined, but that the product DB is not. The products AB and BA are both defined, but they are not of the same size. The products CD and DC are of the same size, but are not equal.

Besides checking the details of computation in this example, you should not continue reading until you are impressed with two facts about matrix multiplication:

1. The product AB of two matrices is defined only when the number of columns of A is equal to the number of rows of B. The product has the same number of rows as A and the same number of columns as B.
2. Matrix multiplication is not commutative. Even if the products AB and BA are both defined and are both the same size, the matrices AB and BA are not necessarily equal.

The fact that, in general, $AB \neq BA$ is the first significant difference between matrix arithmetic and ordinary arithmetic. There are other surprising results in store. We know that if the product of two numbers is zero, at least one of the factors must be zero. This is not true for matrices. The next example illustrates what can happen.

EXAMPLE

Let

$$A = \begin{pmatrix} 4 & 2 \\ -2 & -1 \end{pmatrix} \quad \text{and} \quad B = \begin{pmatrix} 3 & 2.5 \\ -6 & -5 \end{pmatrix}.$$

Then the product AB is the zero matrix

$$AB = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$$

although no entry of either factor is a zero.

Before you give up in despair, it is well to point out that a number of properties of ordinary arithmetic continue to hold true for matrix multiplication. Some of these are listed in the next theorem. For convenience, we state the theorem for square matrices although some of the results hold more generally.

THEOREM 2 Let A, B, and C be any $n \times n$ matrices, and let c be any constant. Then the following properties are all true:

1. AB is an $n \times n$ matrix
2. $A(BC) = (\mathrm{AB})\mathrm{C}$
3. $A(B + C) = AB + AC$
4. $(B + C)A = BA + CA$
5. $c(AB) = (cA)B = A(cB)$
6. There is a unique $n \times n$ matrix I such that $AI = IA = A$ for every $n \times n$ matrix A.

Note: The matrix I of Property (56) is called the *identity matrix* and is defined by the condition:

$$I_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$

That is, the identity matrix has the form

$$I = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ .. & \dots & .. & \dots & 0 \\ 0 & 0 & 0 & \dots & 1 \end{pmatrix}$$

In particular the 2×2 and 3×3 identity matrices have the form $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ and $\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$, respectively.

Proof of Theorem 2 The hardest result to establish is (2). The others are substantially easier and will be left, as is the custom, as exercises.

We let D represent the matrix BC and let $E = AB$. We need to show that the ijth entry of $[A(BC)]$ is the same as the ijth entry of $[(AB)C]$. Now we have

$$[A(BC)]_{ij} = (AD)_{ij} = \sum_{k=1}^{n} a_{ik} d_{kj}$$

where

$$d_{kj} = (BC)_{kj} = \sum_{r=1}^{n} b_{kr} c_{rj}$$

so that

$$[A(BC)]_{ij} = \sum_{k=1}^{n} a_{ik} \sum_{r=1}^{n} b_{kr} c_{rj}$$

but since ordinary arithmetic of real numbers is commutative and associative, we write this double sum as

$$\sum_{k=1}^{n}\sum_{r=1}^{n} a_{ik}(b_{kr}c_{rj}) = \sum_{k=1}^{n}\sum_{r=1}^{n} a_{ik}b_{kr}c_{rj} = \sum_{r=1}^{n}\sum_{k=1}^{n} a_{ik}b_{kr}c_{rj} = \sum_{r=1}^{n}\sum_{k=1}^{n} (a_{ik}b_{kr})c_{rj}$$
$$= \sum_{r=1}^{n} e_{ir}c_{rj} = (EC)_{ij} = [(AB)C]_{ij}$$

where

$$e_{ir} = E_{ir} = (AB)_{ir} = \sum_{k=1}^{n} a_{ik} b_{kr}$$

The fact that the matrices $(AB)C$ and $A(BC)$ are identical means that we can ignore the parentheses and write ABC *to* represent the product. It also means that we may define, unambiguously, positive integral powers of a square matrix. That is, if A is an $n \times n$ matrix, then $A^2 = AA$, $A^3 = AAA$, $A^4 = AAAA$, and so on. ◇

Inverses

In your first studies of algebra, you learned to solve equations of the form

$$ax + c = d$$

where a, c, and d were given numbers and x was an unknown number.

We can form analogous algebraic questions for matrices. Suppose for example that A, C, and D are given $n \times n$ matrices. Does there exist an $n \times n$ matrix X such that

$$AX + C = D \tag{2}$$

If so, how do we compute X?

Since matrices of the same size can be subtracted, Eq. (2) is equivalent to

$$AX = D - C \tag{3}$$

Now suppose there is an $n \times n$ matrix B so that $BA = I$ where I is the $n \times n$ identity matrix. If we multiply each side of Eq. (3) on the left by B, we obtain

$$B(D - C) = B(AX) = (BA)X = IX = \mathrm{X}.$$

The question of solving the matrix Eq. (2) reduces then to finding a matrix B with the stated property.

DEFINITION Let A be an $n \times n$ matrix. Any $n \times n$ matrix B such that $AB = BA = I$ is called an *inverse* of A.

Corollary of the Definition If B is an inverse of A, then A is an inverse of B.

EXAMPLE

The matrix

$$B = \begin{pmatrix} 3 & -4 \\ -5 & 7 \end{pmatrix}$$

is an inverse of the matrix

$$A = \begin{pmatrix} 7 & 4 \\ 5 & 3 \end{pmatrix}$$

You may verify this claim by computing the products AB and BA.

The definition of an inverse does not rule out the possibility that a matrix may have more than one inverse. Our first theorem about inverses shows that this cannot happen; if a matrix has an inverse, then it has a unique one.

THEOREM 3 If B and B' are inverses of the matrix A, then $B = B''$.

Proof of Theorem 3 The proof is a clever, one-line affair:

$$B = BI = B(AB') = (BA)B' = IB' = B'.$$

According to Theorem 3, we may speak of *the* inverse of a matrix. Since the identity matrix I plays the same role in matrix multiplication as the number 1 in the multiplication of numbers, the inverse of a matrix plays the role of the reciprocal. For this reason, the inverse of the matrix A is denoted by A^{-1}. ◇

As an application of the inverse of a matrix, consider the following example.

EXAMPLE

If a Holstein cow is fed x units of grain and y units of hay per day, then she will produce $7x + 4y$ pounds of skim milk and $5x + 3y$ pounds of butterfat per week. How much would you have to feed her to get 41 pounds of milk and 30 pounds of butterfat?

Solution

Let A be the matrix

$$A = \begin{pmatrix} 7 & 4 \\ 5 & 3 \end{pmatrix}$$

and let X be a 2×1 matrix whose components are units of grain and hay, respectively. Then AX is a 2×1 matrix whose components represent the pounds of skim milk and butterfat, respectively. Let **D** be the column vector of desired output

$$D = \begin{pmatrix} 41 \\ 30 \end{pmatrix}$$

Then we are trying to find a vector **X** such that $A\mathbf{X} = \mathbf{D}$. Multiplying each side of this equation on the left by A^{-1} gives the answer: $\mathbf{X} = A^{-1}\mathbf{D}$. Since A^{-1} is

$$A^{-1} = \begin{pmatrix} 3 & -4 \\ -5 & 7 \end{pmatrix}$$

we have

$$X = \begin{pmatrix} 3 & -4 \\ -5 & 7 \end{pmatrix} \begin{pmatrix} 41 \\ 30 \end{pmatrix} = \begin{pmatrix} 3 \\ 5 \end{pmatrix}$$

The farmer should feed the cow 3 units of grain and 5 units of hay.

Existence of Inverses

It is not true that every square matrix has an inverse. Consider, for example, the matrix

$$A = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$$

If B is any 2×2 matrix,

$$B = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

then the product AB has the form

$$AB = \begin{pmatrix} a & 0 \\ c & 0 \end{pmatrix}$$

and no choice of a and c will make this the identity matrix. Thus, A has no inverse.

Clearly, the presence of so many zeros as entries of A has something to do with the lack of an inverse. However, less suspicious-looking matrices may also fail to possess inverses.

EXAMPLE

The matrix C, given by

$$C = \begin{pmatrix} 3 & 6 \\ 2 & 4 \end{pmatrix}$$

does not have an inverse.

Suppose, to the contrary, that there was a matrix B such that $CB = I$. Now the product of C and the first column of the matrix B must give the first column of the identity matrix. If the first column of B looks like

$$\begin{pmatrix} a \\ c \end{pmatrix}$$

then we must have the two equations

$$3a + 6c = 1$$
$$2a + 4c = 0$$

But if $2a + 4c = 0$, then $a + 2c = 0$, so that $3a + 6c = 0$ and cannot equal 1.

The existence of an inverse for a square matrix hinges, then, on the question of whether a certain system of linear algebraic equations has a solution.

The problem of determining whether an inverse exists and, if it does, of computing it is somewhat simplified by the following theorem.

THEOREM 4 If A and B are square matrices of order n and $AB = I$, then BA is also the identity.

Because the proof of this theorem is not elementary, we will not present it. Any standard linear algebra text will contain the proof. For example, see Section 2.3 of David C. Lay, *Linear Algebra and Its Applications*, 4th ed., Boston: Pearson Addison Wesley, 2012.

Suppose, then, that we are given a 2×2 matrix A:

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

According to Theorem 4, A will have an inverse exactly if there is a matrix

$$B = \begin{pmatrix} w & x \\ y & z \end{pmatrix}$$

such that

$$AB = I = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

The matrix equality $AB = I$ translates into a system of four linear equations in four unknowns:

$$\begin{aligned} aw + by &= 1 \\ ax + bz &= 0 \\ cw + dy &= 0 \\ cx + dz &= 1 \end{aligned} \tag{4}$$

The existence of an inverse for a given 3×3 matrix reduces similarly to the existence of a solution of a system of nine linear equations in nine unknowns.

The system of Eq. (4) splits quite naturally into two systems, each containing two linear equations in two unknowns:

$$\begin{aligned} aw + by &= 1 \\ cw + dy &= 0 \end{aligned}$$

and

$$\begin{aligned} ax + bz &= 0 \\ cx + dz &= 1 \end{aligned}$$

Note that the coefficients of the unknown terms on the left-hand sides of these two systems are the same. The systems correspond to the matrix problem of finding column vectors $\mathbf{X_1}$ and $\mathbf{X_2}$ so that $\mathbf{AX_1}$ and $\mathbf{AX_2}$ are the first and second columns of the 2×2 identity matrix.

In general, if A is an $n \times n$ matrix, then A has an inverse if and only if there are column vectors $\mathbf{X_1}, \mathbf{X_2}, ..\mathbf{X_n}$ so that $A\mathbf{X}$ is the ith column of the $n \times n$ identity matrix, $i = 1, 2, \ldots, n$. Thus, the problem of finding the inverse or determining its nonexistence reduces to an algebraic problem: determine the nature or nonexistence of solutions to n systems of linear equations, where each system contains n equations in n unknowns. The coefficients of the unknowns in all the systems are the same; only the constants on the right-hand side change. There is a systematic method for solving this problem—the Gauss-Jordan elimination process. We discuss it in detail in Appendix III.

EXERCISES

1. Consider the three row vectors $\mathbf{u} = (4, 2, 3)$, $\mathbf{v} = (-2, 3, 0)$, and $\mathbf{w} = (-1, 1, 1)$. Compute each of the following:

(a) $2\mathbf{u}$

(b) $-\mathbf{v}$

(c) $3\mathbf{u} - 2\mathbf{v}$

(d) $\mathbf{u} + \mathbf{w}$

(e) $\mathbf{u} - \mathbf{v} + \mathbf{w}$

(f) $4\mathbf{u} - 3\mathbf{v} + 2\mathbf{w}$

2. If $3\mathbf{v} - 2\mathbf{w} = \mathbf{0}$ for a pair of vectors $\mathbf{v}$ and $\mathbf{w}$, what is the relationship between the components of $\mathbf{v}$ and $\mathbf{w}$?

3. Suppose $4\mathbf{u} - 2\mathbf{v} + 3\mathbf{w} = \mathbf{0}$ for three vectors $\mathbf{u}$, $\mathbf{v}$, $\mathbf{w}$. What is the relationship among the components of these vectors?

4. If A is a matrix, then we say $A \geq \mathbf{0}$ if every entry of A is nonnegative.

(a) Define $A \leq \mathbf{0}$ analogously.

(b) Prove that if $A \geq \mathbf{0}$, then $-A \leq \mathbf{0}$.

5. If A and B are matrices of the same size, define $A \geq B$ to mean $A - B \geq \mathbf{0}$. Show that if $A \geq B$ and $B \geq C$, then $A \geq C$

6. Suppose A, B, C and D are matrices whose sizes are 3×4, 5×4, 4×4, and 4×3, respectively. Find the sizes of each of the following:

(a) AC

(b) CB

(c) DA

(d) ADC

(e) $BCDA$

7. Let matrices A, B, and C be given by

$$A = \begin{pmatrix} 2 & 0 & -3 \\ 1 & -1 & 4 \\ 3 & 2 & 1 \end{pmatrix} \quad B = \begin{pmatrix} 5 & 6 \\ -7 & 0 \\ 8 & 2 \end{pmatrix} \quad C = \begin{pmatrix} 0 & 0 \\ 2 & -3 \\ 4 & 1 \end{pmatrix}$$

Compute:

(a) $B + \mathrm{C}$

(b) $B - 2C$

(c) $A(B + C)$

(d) A^2

8. Let $\mathbf{0}$ be a zero matrix, and suppose that $A\mathbf{0}$ is defined. Show that $A\mathbf{0}$ is also a zero matrix.

9. Prove Theorem 1.

10. Prove Theorem 2.

11. Show that the system of equations

$$\begin{aligned} 2x - 3y &= 46 \\ 9x + 7y &= 27 \end{aligned}$$

can be represented in matrix form $A\mathbf{u} = \mathbf{v}$ for a suitably chosen 2×2 matrix A and vectors $\mathbf{u}$ and $\mathbf{v}$.

12. Show that any system of linear equations can be written in the form $A\mathbf{u} = \mathbf{v}$ where A is a suitably chosen matrix and vectors $\mathbf{u}$ and $\mathbf{v}$.

13. A matrix is *invertible* if it has an inverse. Suppose A and B are $n \times n$ invertible matrices.

(a) Show that AB is invertible and that $(AB)^{-1} = B^{-1}A^{-1}$.

(b) Is $A + B$ necessarily invertible?

14. Let A be an $m \times n$ matrix and B an $n \times m$ matrix. Then both products AB and BA are square matrices.

(a) If $m > n$, show that AB has no inverse.

(b) If $m > n$, can BA have an inverse?

15. Suppose A is an invertible $n \times n$ matrix with inverse B. Find, where possible, inverses of the following matrices:

(a) A^2

(b) A^3

(c) $2A$

(d) $-A^7$

(e) $(A^{-1})^2$

(f) ABA

16. Let A and B be matrices of the same size and define an operation $A \times B$ by $(A \times B)_{ij} = A_{ij}B_{ij}$; that is, multiply together corresponding entries. Show that this operation is commutative and associative. Is there an "identity" element? Which matrices have "inverses" under this operation? Can you find any applications for this operation?

Solving Systems of Equations

APPENDIX III

The equation

$$7x_1 + 2x_2 = 5 \tag{1}$$

is an example of a *linear equation in two unknowns* x_1 and x_2. This equation is true for some values of the unknowns (for example, $x_1 = 1$, $x_2 = -1$), but false for other values (for example, $x_1 = 2$, $x_2 = 1$). The set of all vectors

$$\mathbf{X} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

for which the equation is true is called the *solution set* of the equation. Any element of this set is called a *solution* of the equation.

It is easy to check that every solution of Eq. (1) is a vector of the form

$$\begin{pmatrix} \alpha \\ \dfrac{5 - 7\alpha}{2} \end{pmatrix}$$

where α can be any real number. Conversely, every vector of this form is a solution of Eq. (1).

The general linear equation in two unknowns has the form

$$a_1x_1 + a_2x_2 = b \tag{2}$$

where a_1, a_2, and b are given constants and a_1 and a_2 are not both zero.

The *general linear equation in* n *unknowns* $x_1, x_2, \ldots, x_n$, is an equation of the form

$$a_1x_1 + a_2x_2, + \cdots + a_nx_n = b \tag{3}$$

where the $a_1, a_2, \ldots, a_n$, and b are given constants and at least one of the a_i's is nonzero. The set of all vectors

$$\mathbf{X} = \begin{pmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{pmatrix} \tag{4}$$

for which Eq. (3) is true is called the *solution set* of (3).

Since a linear equation in two unknowns is the equation of a straight line in the plane, a vector

$$\mathbf{X} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

is a solution if and only if the point (x_1, x_2) lies on the line.

A *system of* m *linear equations in* n *unknowns* is a collection of linear equations

$$\begin{aligned} &a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1 \\ &a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = b_2 \\ &\cdots \\ &\cdots \\ &\cdots \\ &a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n = b_m \end{aligned} \tag{5}$$

where the constants a_{ij} and b_i are given. Note that a_{ij} is the coefficient of x_j in the ith equation.

The *solution set of a system* is defined to be the intersection of the solution sets of the individual equations—that is, a vector

$$\mathbf{X} = \begin{pmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{pmatrix}$$

is a solution of the system (5) if and only if it is a solution of each equation.

In the case of two equations in two unknowns,

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 &= b_1 \\ a_{21}x_1 + a_{22}x_2 &= b_2 \end{aligned} \tag{6}$$

each equation represents a straight line in the plane, so the solution set corresponds to the set of all points lying on both lines. There are three possibilities for the intersection of two lines in the plane: the lines intersect in a single point, the lines are parallel and do not intersect at all, or the lines are coincident.

These three cases are illustrated, respectively, by the examples

$$\begin{aligned} 7x_1 + 2x_2 &= 5 \\ 4x_1 - 3x_2 &= 7 \end{aligned} \tag{7}$$

$$\begin{aligned} 7x_1 + 2x_2 &= 5 \\ 14x_1 + 4x_2 &= 7 \end{aligned} \tag{8}$$

$$\begin{aligned} 7x_1 + 2x_2 &= 5 \\ 14x_1 + 4x_2 &= 10 \end{aligned} \tag{9}$$

The solution set of system (7) consists of the single vector

$$\begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

The solution set of system (8) is empty, and the solution of system (9) is, again, any vector of the form $\mathbf{X} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} \alpha \\ \dfrac{5 - 7\alpha}{2} \end{pmatrix}$, α is arbitrary.

The equations of a general linear system are the equations of "hyperplanes" in n-dimensional space and the solution sets correspond to the points lying on the intersection of these hyperplanes. Although these intersections may take many forms, there are essentially the same three possibilities as for the system of two equations in two unknowns:

(i) Exactly one solution

(ii) No solutions

(iii) Infinitely many solutions

The main purpose of this appendix is to describe a systematic procedure for obtaining the solution set of a system of linear equations. The basic idea is to replace the original system by a sequence of equivalent, progressively simpler systems.

DEFINITION Two systems of equations are called *equivalent* if they have the same solution set—that is, if every solution of either one is a solution of the other.

EXAMPLE

The systems

$$\begin{aligned} 7x_1 + 2x_2 &= 5 \\ 4x_1 - 3x_2 &= 7 \end{aligned} \tag{7}$$

and

$$\begin{aligned} 13x_1 + 9x_2 &= 4 \\ -4x_1 + 7x_2 &= -11 \end{aligned} \tag{10}$$

are equivalent systems.

THEOREM 1 If the positions of any two equations in a system are interchanged to form a new system, then the new system is equivalent to the original system.

EXAMPLE

The systems

$$\begin{aligned} 11x_1 + 12x_2 - 7x_3 &= 8 \\ 3x_1 + 2x_2 + 9x_3 &= 7 \\ x_1 - x_2 + x_3 &= 4 \end{aligned} \tag{11}$$

and

$$\begin{aligned} x_1 - x_2 + x_3 &= 4 \\ 3x_1 + 2x_2 + 9x_3 &= 7 \\ 11x_1 + 12x_2 - 7x_3 &= 8 \end{aligned} \tag{12}$$

are equivalent.

Proof of Theorem 1 Let X_i be the solution set of the ith equation of the original system. Suppose that equations j and k are interchanged. Then the solution set of the original system is

$$X_1 \cap X_2 \cap \ldots X_{j-1} \cap X_j \cap X_{j+1} \cap \ldots \cap X_{k-1} \cap X_k \cap X_{k+1} \cap \ldots \cap X_n$$

and the solution set of the system after interchanging is

$$X_1 \cap X_2 \cap \ldots X_{j-1} \cap X_k \cap X_{j+1} \cap \ldots \cap X_{k-1} \cap X_j \cap X_{k+1} \cap \ldots \cap X_n$$

Since intersection of sets is a commutative operation, the two solution sets are the same. ◇

THEOREM 2 If an equation of a given linear system is replaced by a nonzero multiple of itself plus a multiple of another equation of the system to obtain a new system, then the new system is equivalent to the original system.

EXAMPLE

Suppose the second equation of system (11) is replaced by (-1) times the second equation plus 3 times the first equation:

$$3x_1 + 2x_2 + 9x_3 = 7$$

is replaced by

$$(-3x_1 - 2x_2 - 9x_3 = -7) + (3x_1 - 3x_2 + 3x_3 = 12) = -5x_2 - 6x_3 = 5$$

The new system is

$$\begin{aligned} 11x_1 + 12x_2 - 7x_3 &= 8 \\ 3x_1 + 2x_2 + 9x_3 &= 7 \\ -5x_2 - 6x_3 &= 5 \end{aligned} \tag{13}$$

By Theorem 2, Eqs. (11) and (13) are equivalent. Note that system (13) is "simpler" in the sense that we have eliminated one of the unknowns in one of the equations.

Proof of Theorem 2 If the ith equation of the original system (5) is replaced by c times the ith equation plus d times the jth equation, then the new system is

$$\begin{aligned} &a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1 \\ &\cdots \\ &c(a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{in}x_n) \\ &\quad + d(a_{j1}x_1 + a_{j2}x_2 + \cdots + a_{jn}x_n) = cb_i + db_j \\ &\cdots \\ &a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n = b_m \end{aligned} \tag{14}$$

Now, every vector $\mathbf{X}$ which satisfies the equations of (5) will also satisfy the equations of (14). On the other hand, system (5) can be obtained from system (14) by a similar operation: replace the ith equation of (14) by $(1/c)$ times the ith equation plus $(-d/c)$ times the jth equation. Thus, every vector satisfying the equations of (14) will also satisfy the equations of (5). We have seen that the solution set of each system is a subset of the other. Hence, the solution sets are equal. ◇

The method of solution we shall describe is called the Gauss-Jordan elimination procedure. It consists of a sequence of operations using Theorem 1 and Theorem 2 to obtain new, equivalent systems that eliminate, at each step, at least one unknown in one of the equations. To be more precise, we use Theorem 2 to eliminate x_1 from every equation except the first, then use Theorem 2 to eliminate x_2 from every equation except the second, and so on. Eventually, we obtain a system whose solution set can be determined by inspection. We shall illustrate the procedure with several examples.

EXAMPLE

Consider system (7)

$$\begin{aligned} 7x_1 + 2x_2 &= 5 \\ 4x_1 - 3x_2 &= 7 \end{aligned} \tag{7}$$

Step 1. Replace the first equation with (1/7) times the first equation. We obtain

$$\begin{aligned} x_1 + 2/7\, x_2 &= 5/7 \\ 4x_1 - 3x_2 &= 7 \end{aligned} \tag{15}$$

Step 2. In system (15), replace the second equation with the second equation plus (-4) times the first equation. The result is

$$\begin{aligned} x_1 + 2/7 x_2 &= 5/7 \\ -29/7 x_2 &= 29/7 \end{aligned} \tag{16}$$

Step 3. Replace the second equation with $(-7/29)$ times itself:

$$\begin{aligned} x_1 + \frac{2}{7} x_2 &= 5 \\ x_2 &= -1 \end{aligned} \tag{17}$$

Step 4. Replace the first equation with the first equation plus $(-2/7)$ times the second equation:

$$\begin{aligned} x_1 + \ &= 1 \\ x_2 &= -1 \end{aligned} \tag{18}$$

Now the solution set can be read off from the equations of system (18). It consists of the unique vector

$$\mathbf{X} = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

As a second example, consider the system

$$\begin{aligned} x_2 - 2x_3 &= 0 \\ 2x_1 + x_2 - 4x_3 &= 6 \\ x_1 + x_2 + x_3 &= 3 \end{aligned} \tag{19}$$

Step 1. Interchange the first and third equations:

$$\begin{aligned} x_1 + x_2 + x_3 &= 3 \\ 2x_1 + x_2 - 4x_3 &= 6 \\ x_2 - 2x_3 &= 0 \end{aligned} \tag{20}$$

Step 2. Use the first equation to eliminate x_1 in the other equations. Since x_1 is already missing in the third equation, we only have to work on the second equation. Replace it by the second equation plus (-2) times the first equation:

$$\begin{aligned} x_1 + x_2 + x_3 &= 3 \\ -x_2 + 2x_3 &= 0 \\ x_2 - 2x_3 &= 0 \end{aligned} \tag{21}$$

Step 3. Use the second equation to eliminate x_2 from the other equations:
   a. Replace the second equation with (-1) times the second.
   b. Replace the first equation with the first equation plus (-1) times the second.
   c. Replace the third equation with the third plus (-1) times the second. The result is

$$\begin{aligned} x_1 + 3x_3 &= 3 \\ x_2 - 2x_3 &= 0 \\ 0 &= 0 \end{aligned} \tag{22}$$

From Eq. (22), we see that we can assign any value to x_3 and then compute x_1 and x_2 from the first two equations. For example, if $x_3 = 0$, then $x_1 = 3$ and $x_2 = 0$; if $x_3 = 1$, then $x_1 = 0$ and $x_2 = 2$. Thus, the vectors $\mathbf{X_1} = \begin{pmatrix} 3 \\ 0 \\ 0 \end{pmatrix}$ and $\mathbf{X}_2 = \begin{pmatrix} 0 \\ 2 \\ 1 \end{pmatrix}$ are solutions to the original system. The general solution can be described by letting x_3 take on any arbitrary value α. Then $x_1 = 3 - 3\alpha$ and $x_2 = 2\alpha$. Thus, the solution set is the set of all vectors of the form $\mathbf{X} = \begin{pmatrix} \mathbf{3} - \mathbf{3}\alpha \\ \mathbf{2}\alpha \\ \alpha \end{pmatrix}$, α arbitrary.

If the right-hand side of Eqs. (19) had been replaced by the constant vector

$$\mathbf{b} = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}$$

then we would have obtained the system

$$\begin{aligned} x_2 - 2x_3 &= 1 \\ 2x_1 + x_2 + 4x_3 &= 2 \\ x_1 + x_2 + x_3 &= 3 \end{aligned} \tag{19′}$$

The procedure to solve this system is the same as that for system (19); the operations we use in the Gauss-Jordan process are dictated only by the coefficients of the unknowns $x_1, x_2, \ldots x_n$ and are independent of the constants on the right-hand sides of the equations.

After completing Steps 1, 2, and 3, we would arrive at

$$\begin{aligned} x_1 + 3x_3 &= -1 \\ x_2 - 2x_3 &= 4 \\ 0 &= -3 \end{aligned} \tag{22′}$$

The system (22′) has no solution, since the third equation is not true for any choice of x_1, x_2, and x_3. Thus, the equivalent system (19′) has no solution. Geometrically, the equations of (19′) represent three planes in three-dimensional space that have no common point. The three planes of system (19) intersect along a line.

We can simplify the Gauss-Jordan elimination procedure somewhat by adopting matrix notation. The original system of linear Eq. (5) can be represented by the matrix equation

$$A\mathbf{X} = \mathbf{B}$$

where A is the $m \times n$ matrix of coefficients

$$A = \begin{pmatrix} a_{11} & a_{12} & . & . & . & a_{1n} \\ a_{21} & a_{22} & & & & a_{2n1} \\ . & & & & & \\ . & & & & & \\ . & & & & & \\ a_{m1} & a_{m2} & . & . & . & a_{mn} \end{pmatrix}$$

and $\mathbf{X}$ and $\mathbf{B}$ are the column vectors

$$\mathbf{X} = \begin{pmatrix} x_1 \\ x_2 \\ . \\ . \\ . \\ x_n \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} b_1 \\ b_2 \\ . \\ . \\ . \\ b_m \end{pmatrix}$$

In the Gauss-Jordan process, the entries of A and $\mathbf{B}$ will change after each operation. We can keep track of these changes by considering an $m \times (n+1)$ *augmented matrix*

$$(A|\mathbf{B}) = \left(\begin{array}{cccccc|c} a_{11} & a_{12} & . & . & . & a_{1n} & b_1 \\ a_{21} & a_{21} & . & . & . & a_{2n} & b_2 \\ . & & & & & & \\ . & & & & & & \\ . & & & & & & \\ a_{m1} & a_{m2} & . & . & . & a_{mn} & b_m \end{array}\right)$$

Then the operations of Theorems 1 and 2 can be interpreted as operations on the rows of the augmented matrix:

a. Interchange two rows of $(A|\mathbf{B})$.
b. Replace one row of $(A|\mathbf{B})$ with the sum of a nonzero multiple of that row and a multiple of another row.

EXAMPLE

The system

$$2x_1 + 10x_2 + 6x_3 = 14$$
$$4x_1 + 22x_2 - 8x_3 = 12$$

the augmented matrix

$$\left(\begin{array}{ccc|c} 2 & 10 & 6 & 14 \\ 4 & 22 & -8 & 12 \end{array}\right)$$

To solve the system, we carry out the operations of the Gauss-Jordan procedure on the rows of the augmented matrix.

Step 1. Replace the first row with $(1/2)$ the first row. The result is

$$\left(\begin{array}{ccc|c} 1 & 5 & 3 & 7 \\ 4 & 22 & -8 & 12 \end{array}\right)$$

Step 2. Replace the second row with the second row plus (-4) times the first row. The result is

$$\left(\begin{array}{ccc|c} 1 & 5 & 3 & 7 \\ 0 & 2 & -20 & -16 \end{array}\right)$$

Step 3. Replace the second row with $(1/2)$ times the second row:

$$\left(\begin{array}{ccc|c} 1 & 5 & 3 & 7 \\ 0 & 1 & -10 & -8 \end{array}\right)$$

Step 4. Replace the first row with the first row plus (-5) times the second row:

$$\left(\begin{array}{ccc|c} 1 & 0 & 53 & 47 \\ 0 & 1 & -10 & -8 \end{array}\right)$$

This augmented matrix represents the system

$$x_1 + 53x_3 = 47$$
$$x_2 - 10x_3 = -8$$

which has as its solution set the set of vectors of the form

$$\mathbf{X} = \begin{pmatrix} 47 - 53\alpha \\ -8 + 10\alpha \end{pmatrix}, \alpha \text{ arbitrary.}$$

With these examples in mind, we may describe the Gauss-Jordan procedure more explicitly:

1. Interchange equations (or rows of the augmented matrix) so that the first equation has a nonzero coefficient of x_1.
2. Replace the first equation with the first equation multiplied by the reciprocal of the coefficient of x_1.
3. Use the new first equation to eliminate x_1 in every other equation. The ith equation is replaced with the ith equation plus $(-a_{il})$ times the first equation $(i = 2, 3, \ldots, m)$.
4. Let j be the smallest number such that x_j occurs with some nonzero coefficient in some equation other than the first. Interchange equations so that the new second equation has a nonzero coefficient of x_j.
5. Replace the second equation with the second equation multiplied by the reciprocal of the coefficient of x_j.
6. Use the new second equation to eliminate x_j in all equations (including the first equation) except the second. Follow the procedure of the third step to do this.
7. Let k be the smallest number for which x_k appears with a nonzero coefficient in some equation other than the first two. Make this the new third equation and use it to eliminate x_k in every equation except the third.
8. Continue in this manner until further simplification is not possible. Read off the solution set from the resulting system.

Computing the Inverse of a Square Matrix

The Gauss-Jordan procedure can be used for any system of m linear equations in n unknowns. In this section, we apply the procedure to the problem of determining the inverse of a square matrix.

Suppose we wish to find the inverse of the matrix

$$A = \begin{pmatrix} 7 & 4 \\ 5 & 3 \end{pmatrix}$$

The discussion in Appendix II shows that we need to find vectors $\mathbf{X_1}$ and $\mathbf{X_2}$, so that

$$A\mathbf{X_1} = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \quad \text{and} \quad A\mathbf{X}_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

That is, we need to solve the systems

$$\begin{aligned} 7x_1 + 4x_2 &= 1 \\ 7x_1 + 4x_2 &= 0 \end{aligned}$$

and

$$5x_1 + 3x_2 = 0$$
$$5x_1 + 3x_2 = 1$$

There are two approaches we may take.

Approach I Solve the more general system

$$7x_1 + 4x_2 = a$$
$$5x_1 + 3x_2 = b$$

and then find $\mathbf{X_1}$ by letting $a = 1$ and $b = 0$, and find $\mathbf{X_2}$ by letting $a = 0$ and $b = 1$.

The augmented matrix is

$$\left(\begin{array}{cc|c} 7 & 4 & a \\ 5 & 3 & b \end{array}\right)$$

The steps in the Gauss-Jordan procedure are:

Step 1. Divide row 1 by 7:

$$\left(\begin{array}{cc|c} 1 & \frac{4}{7} & \frac{a}{7} \\ 5 & 3 & b \end{array}\right)$$

Step 2. Replace row 2 with row $2 - (5)$ (row 1):

$$\left(\begin{array}{cc|c} 1 & \frac{4}{7} & \frac{a}{7} \\ 0 & \frac{1}{7} & \frac{7b - 5a}{7} \end{array}\right)$$

Step 3. Replace row 2 with 7 (row 2):

$$\left(\begin{array}{cc|c} 1 & \frac{4}{7} & \frac{a}{7} \\ 0 & 1 & 7b - 5a \end{array}\right)$$

Step 4. Replace row 1 with row $1 - (4/7)$ (row 2):

$$\left(\begin{array}{cc|c} 1 & 0 & 3a - 4b \\ 0 & 1 & 7b - 5a \end{array}\right)$$

Letting $a = 1$, $b = 0$, we obtain

$$\mathbf{X_1} = \begin{pmatrix} 3 \\ -5 \end{pmatrix}$$

Letting $a = 0$, $b = 1$ produces

$$\mathbf{X_2} = \begin{pmatrix} -4 \\ 7 \end{pmatrix}$$

Thus, the inverse of A is

$$A^{-1} = (\mathbf{X_1}, \mathbf{X_2}) = \begin{pmatrix} 3 & -4 \\ -5 & 7 \end{pmatrix}$$

Approach II Solve the two systems simultaneously by using a doubly augmented matrix

$$\left(\begin{array}{cc|cc} 7 & 4 & 1 & 0 \\ 5 & 3 & 0 & 1 \end{array}\right)$$

Since the steps for solution are the same, we merely note the augmented matrices:

After Step 1: $\left(\begin{array}{cc|cc} 1 & \frac{4}{7} & \frac{1}{7} & 0 \\ 5 & 3 & 0 & 1 \end{array}\right)$

After Step 2: $\left(\begin{array}{cc|cc} 1 & \frac{4}{7} & \frac{1}{7} & 0 \\ 0 & \frac{1}{7} & \frac{-5}{7} & 1 \end{array}\right)$

After Step 3: $\left(\begin{array}{cc|cc} 1 & \frac{4}{7} & \frac{1}{7} & 0 \\ 0 & 1 & -5 & 7 \end{array}\right)$

After Step 4: $\left(\begin{array}{cc|cc} 1 & 0 & 3 & -4 \\ 0 & 1 & -5 & 7 \end{array}\right)$

Note that in this approach, we begin with $(A|I)$ and end up with $(I|A^{-1})$.

No matter which approach is used, it is important that the operation used at each step be applied to all the coefficients in the indicated row of the augmented matrix.

A FINAL EXAMPLE

Consider the system

$$\begin{aligned} x_1 + x_2 + x_3 &= a \\ 2x_1 - 4x_2 + 7x_3 &= b \\ -x_1 + 5x_2 - 6x_3 &= c \end{aligned}$$

which has the augmented matrix

$$\left(\begin{array}{ccc|c} 1 & 1 & 1 & a \\ 2 & -4 & 7 & b \\ -1 & 5 & -6 & c \end{array}\right)$$

The reader should verify that the following steps, in the indicated order—

1. Replace row 2 – 2 row 1
   replace row 3 with row 3 + row 1
2. Replace row 2 with $(-1/6)$ row 2
   replace row 1 with row 1 + (-1) row 2
   replace row 3 with row 3 – 6 (row 2)
   —yield the augmented matrix

$$\left(\begin{array}{ccc|c} 1 & 0 & \frac{11}{6} & \frac{8a-b}{6} \\ 0 & 1 & \frac{-5}{6} & 2a-b \\ 0 & 0 & 0 & -a+b+c \end{array}\right)$$

The corresponding system of equations has a solution if and only if $-a+b+c=0$. In particular, if $a=1$, $b=0$, $c=0$, then $-a+b+c=-1\neq 0$. Thus, if A is the matrix

$$A=\begin{pmatrix} 1 & 1 & 1 \\ 2 & -4 & 7 \\ -1 & 5 & -6 \end{pmatrix}$$

then it is impossible to find a vector $\mathbf{X}$ such that

$$A\mathbf{X}=\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$$

Hence, the matrix A does not have an inverse.

EXERCISES

1. Solve the system

$$x_1 + 4x_2 + 3x_3 = 1$$
$$-3x_2 - 2x_3 = 2$$
$$-7x_2 - 5x_3 = 4$$

2. Solve the system

$$x_1 - 2x_2 - 3x_3 = 2$$
$$x_1 - 4x_2 - 13x_3 = 14$$
$$-3x_1 + 5x_2 + 4x_3 = 0$$

3. Solve the system

$$x + 2y + z = 3$$
$$3x + 6y + 11z = 8$$
$$-2x - 4y + 4z = 9$$

4. Solve the system

$$8x - 8y + 2u + 4v + 2w = -14$$
$$4x + 2y - 2u - v + 7w = 29$$
$$x + 4y + 3u + 5v + 7w = 2$$

for x, y, v in terms of u and w. (No promises that the arithmetic will be simple.)

5. The Gauss-Jordan procedure works even if the number of equations equals or exceeds the number of unknowns. Solve the following system for w and x in terms of y and z:

$$w + 2x + 3y + 4z = 10$$
$$2w - x + y - z = 1$$
$$3w + x + 4y + 3z = 11$$
$$-2w + 6x + 4y + 10z = 18$$

6. The system

$$x + y + z + w = 8$$
$$x - 2y + 4z = -1$$
$$2x - y + 5z + w = 6$$

is inconsistent—that is, it has no solutions (add the first two equations together and compare the result with the third equation). Try to solve for x in terms of y, z, w using the Gauss-Jordan procedure, and discuss what happens.

7. Construct a flow chart for the Gauss-Jordan procedure.

8. Show that when the Gauss-Jordan procedure is used to find an inverse of a square matrix, the result is either that there is no inverse or there is a unique inverse—that is, show why there cannot be infinitely many solutions of the corresponding system of linear equations.

9. Let $\mathbf{X_1}$ and $\mathbf{X_2}$ be any two solutions of the matrix equation $A\mathbf{X} = 0$ where 0 is a zero matrix.

(a) Show that $\mathbf{X_1} + \mathbf{X_2}$ is also a solution.

(b) Show that $c\mathbf{X_1}$ is a solution where c is any constant.

(c) Show that $A\mathbf{X} = 0$ always has at least one solution, and that if it has two distinct solutions, then it must have infinitely many distinct solutions.

10. Let $\mathbf{X}^*$ be any solution of $A\mathbf{X} = B$ where A and B are given matrices, and let $\mathbf{X_0}$ be a solution of $A\mathbf{X} = 0$.

(a) Show that $\mathbf{X}^* + \mathbf{X_0}$ is a solution of $A\mathbf{X} = B$.

(b) Show that every solution of $A\mathbf{X} = B$ can be written in the form $\mathbf{X} = \mathbf{X}^* + \mathbf{X_1}$ where $\mathbf{X_1}$ is some solution of $A\mathbf{X} = 0$.

11. Find, if possible, inverses of the following matrices:

$$A = \begin{pmatrix} 3 & 1 \\ 11 & 4 \end{pmatrix} \quad B = \begin{pmatrix} 5 & 6 \\ 4 & 5 \end{pmatrix} \quad C = \begin{pmatrix} 8 & 4 \\ 6 & 3 \end{pmatrix}$$

$$D = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \quad E = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix} \quad F = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$$

12. Let A be the 2×2 matrix

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

(a) Show that A has an inverse if and only if $ad - bc \neq 0$.

(b) Determine A^{-1} if $ad - bc \neq 0$.

13. Let B be an arbitrary 3×3 matrix. Find necessary and sufficient conditions on the entries of B for the matrix to have an inverse.

14. The Otter Creek Manufacturing Company produces two kinds of skis, "Premium" and "Quality." From x pounds of wood and y pounds of plastic, it can produce $7x + 4y$ pairs of Premium skis and $5x + 3y$ pairs of Quality skis. If there is a demand for 5,600 pairs of Premiums and 4,100 pairs of Quality skis, how much wood and plastic should the company order?

Functions of Two Variables

APPENDIX IV

Let S be a subset of the (x, y)-plane. A relationship that assigns a unique number to each point of S is called a *real-valued function of two variables*. The domain of such a function is a set of ordered pairs (x, y) of real numbers, and the range is a subset of the reals. We may denote such a function by the letters customarily reserved for functions: f, g, h, F, G, H, φ, θ, ...

We write

$$z = f(x, y)$$

to denote that f assigns the number z to the ordered pair (x, y).

EXAMPLE 1

Consider the function $f(x, y) = x^2 + y^4$. We then have $f(9, 2) = 9^2 + 2^4 = 81 + 16 = 97$, $f(-9, 2) = 97$, and $f(-7, 0) = 49$. This function is defined for all values of x and y, so its domain is the entire plane. Since $x^2 + y^4$ is the sum of two nonnegative numbers, no negative numbers can be in the range. On the other hand, if z is any nonnegative real number, then $f(z^{1/2}, 0) = z$. Thus, the range of f is the set of all nonnegative real numbers. Since $f(9, 2) = f(-9, 2) = f(9, -2) = f(-9, -2)$, the function is not one-to-one.

EXAMPLE 2

Let f be the function given by $f(x, y) = \frac{1}{y\sqrt{x}}$. Then we have $f(4, 3) = 1/6$, $f(9, -1/3) = -1$, while $f(-5, 2)$ and $f(2, 0)$ are undefined. Now f is defined whenever x is positive and y is nonzero. Thus, the domain of f is the open right half-plane excluding the x-axis—that is, $S = \{(x, y) : x > 0, y \neq 0\}$. The range *of f* consists of all real numbers except 0, for if $z \neq 0$, then $f(1, 1/z) = z$.

CONTINUITY

The definition of continuity for a function of two variables is much the same as that for a function of one variable. The basic idea is exactly the same: small changes in domain values

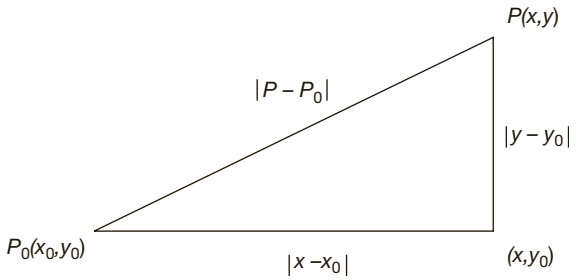


Fig. IV.1 The distance $|P - P_0|$ from P_0 to P is the length of the hypotenuse of a right triangle with sides of length $|x - x_0|$ and $|y - y_0|$.

Table IV.1

$f(x, y)$	$f_x(x, y)$	$f_y(x, y)$	$f_x(9, 2)$	$f_y(9, 2)$	$f_x(-9, 2)$	$f_y(-9, 2)$
$x^2 + y^4$	$2x$	$4y^3$	18	32	–18	32
x^2y	$2xy$	x^2	36	81	–36	81
$\frac{1}{y\sqrt{x}}$	$\frac{-1}{2y\sqrt{x^3}}$	$\frac{-1}{y^2\sqrt{x}}$	$\frac{-1}{108}$	$\frac{-1}{12}$	undefined	undefined

yield relatively small changes in range values. We obtain a more precise definition by using the ε-δ approach.

DEFINITION A function f of one variable is *continuous* at x_0 if for every positive number ε there is a positive number δ such that $|f(x) - f(x_0)| < \varepsilon$ whenever $|x - x_0| < \delta$.

DEFINITION A function f of two variables is *continuous* at $P_o = (x_0, y_0)$ if for every positive number ε there is a positive number δ such that $|f(P) - f(P_o)| < \varepsilon$ whenever $|P - P_o| < \delta$. Here P is a point (x, y) in the plane and $|P - P_o|$ is the Euclidean distance $\sqrt{(x - x_0)^2 + (y - y_0)^2}$ between points in the plane, derived from the Pythagorean Theorem (see Fig. IV.1).

Partial Derivatives

Recall that a derivative of a function of one variable $y = f(x)$ is a measure of the rate of change of the "dependent" variable y with respect to changes in the "independent" variable x. For a function of two variables, $f(x, y) = z$, we have a dependent variable z, and two independent variables, x and y. We can measure rates of change of z with respect to x and with respect to y.

Computationally, partial derivatives are easy to find. The partial derivative of f with respect to x is denoted by $\partial f/\partial x$ or $\partial z/\partial x$ or f_x. To compute f_x, simply pretend that y is a constant and carry out ordinary differentiation with respect to x on the formula for $f(x, y)$. The partial derivative of f with respect to y, denoted by $\partial f/\partial y$ or by $\partial z/\partial y$ or f_y, is computed in an analogous fashion. Some examples are provided in Table IV.1.

The formal definitions of the partial derivatives look like this:

$$f_x(x_0, y_0) = \lim_{h \to 0} \frac{f(x_0 + h, y) - f(x_0, y_0)}{h}$$

if the limit exists, and

$$f_y(x_0, y_0) = \lim_{h \to 0} \frac{f(x_0, y_0 + h) - f(x_0, y_0)}{h}$$

if this limit exists.

Higher-Order Derivatives

Since f_x and f_y are also functions of two variables, we can compute the partial derivatives of each of these. This leads to four *second-order partial derivatives*: f_{xx}, f_{xy}, f_{yx}, and f_{yy}.

For the example $f(x, y) = x^2 + y^4$, we have

$$f_{xx}(x, y) = 2, f_{xy}(x, y) = 0 = f_{yx}(x, y), f_{yy}(x, y) = 12y^2$$

There is nothing to stop us now from computing partial derivatives of higher and higher orders. For instance, the expression f_{xyxx} would denote the function of two variables obtained by differentiating f first with respect to x, then with respect to y, and then twice more with respect to x.

This process might break down, however—for example, the function $f(x, y) = x^{3/2}y$ is defined and continuous everywhere, and so are the functions

$$f_x(x, y) = \frac{3}{2}x^{1/2}y, \quad f_y(x, y) = x^{3/2}$$

$$f_{xy}(x, y) = \frac{3}{2}x^{1/2} = f_{yx}(x, y)$$

but the second-order partial derivative f_{xx} is undefined whenever $x = 0$.

The *graph* of a function of two variables with domain S is the set

$$G_f = \{(x, y, z) : (x, y) \text{ is in } S \text{ and } z = f(x, y)\}$$

The set G_f represents a two-dimensional surface in three-dimensional space. The higher the order of partial derivatives that exist for f, the "smoother" this surface will be.

It is a remarkable result that if f_{xy} and f_{yx} are both continuous functions, then they are equal. Check this for the functions in Table IV.1.

EXERCISES

1. Determine f_x and f_y where $f(x, y)$ is given by
- **a.** $\sin x \cos y$
- **b.** $x^2 + x \sin(x + y)$
- **c.** e^{x+2y}
- **d.** $\log(x - y^2)$

2. Check whether $fxy = frx$ where f is given by
- **a.** $x^2y + y^2x + xy$
- **b.** $cos\,(x^2 - y^2)$
- **c.** $(x+y)^{-1}$

3. Compute $f_{xx} + f_{yy}$ for
- **a.** $f(x, y) = x^3 - 3xy^2$
- **b.** $f(x, y) = \log(x^2 + y^2)$

4. Find the domain and range of each function of two variables given in this appendix and the exercises.

APPENDIX V Differential Equations

In this text we deal with a number of differential equations of a fairly simple type. The following paragraphs will serve as an introduction for those who have not studied this topic before.

By a *first-order differential equation*, we will mean an equation of the form

$$\frac{dy}{dx} = F(x, y) \tag{1}$$

where F is a given function of two variables defined on some region R of the (x, y)-plane. Eq. (1) asserts that y is a differentiable function of x over some interval $[a, b]$ and that the derivative satisfies (1) for all values of x in that interval. A more precise statement would be that there is a function $y = f(x)$ such that

A. f is a differentiable function of x on $[a, b]$

B. The graph $\{(x, y) : y = f(x),\ a \le x \le b\}$ is contained in R

C. $f'(x^*) = F(x^*, f(x^*))$ for all x^* in $[a, b]$

EXAMPLE 1

Consider the differential equation

$$\frac{dy}{dx} = 3x^2y + \frac{y}{2\sqrt{x}} \tag{2}$$

Here the function $F(x, y)$ is given by $F(x, y) = 3x^2y + \frac{y}{2\sqrt{x}}$.

This function is defined everywhere on the region R of the plane consisting of all points with positive first coordinate, the "open right half-plane." If a and b are any two positive numbers with $a < b$, then the function

$$y = f(x) = 8e^{x^3 + \sqrt{x}} \tag{3}$$

satisfies (A), (B), and (C). You should verify this (immediately).

Note that the function

$$y = f(x) = 8e^{x^3 + \sqrt{x} - 7} \tag{4}$$

also satisfies (A), (B), and (C).

Any function fulfilling the conditions of (A), (B), and (C) is called a *solution* of the differential equation (1). Thus the functions described in Eqs. (3) and (4) are two different solutions of the differential equation (2).

It should be pointed out at once that not every differential equation has a solution. Some continuity properties about F and its partial derivatives usually need to be satisfied to guarantee that a solution exists. Example 1 illustrates the fact that a differential equation may have more than one solution. In fact, the differential equation (2) has infinitely many solutions (list some). There is a uniqueness result, however. If the function F satisfies certain continuity properties and (x_0, y_0) is any point in R, then there is a unique solution of the differential equation $dyfdx = F(x, y)$ whose graph passes through (x_0, y_0).

EXAMPLE 1 (continued)

The function $y = f(x) = 8e^{x^3+\sqrt{x}}$ is the unique solution of (2) whose graph goes through the point $(4, 8e^{66})$.

The exact statement of the basic existence and uniqueness theorem for first-order differential equations follows. Its proof can be found in many standard textbooks on differential equations; two of these are listed in the References on the text website.

THEOREM Suppose the functions F and F_y are continuous at all points of the region R of the (x, y)-plane. Let (x_0, y_0) be any point of R. Then there is an interval I of real numbers containing x_0, and

1. There is a unique function $y = f(x)$ that is a solution of the differential equation $dy/dx = F(x, y)$ for which $f(x_0) = y_0$.
2. The solution exists for all values of x for which the points $(x, f(x))$ lie in R.
3. The solution f varies continuously with the choice of x, x_0, and y_0.

In this appendix, we will discuss techniques for discovering solutions to differential equations when the function $F(x, y)$ takes on one of three very special forms.

Case 1 (Classic Integration) Suppose $F(x, y) = g(x)$; that is, F is a function only of x. In this case, the differential equation has the form

$$\frac{dy}{dx} = g(x) \tag{5}$$

To solve this equation we need only find a function whose derivative is $g(x)$. This is the classic integration problem of elementary calculus. The solution of (5) as the form

$$y = \int g(x)\,dx \tag{6}$$

When the indefinite integration of (6) is carried out, we are left with a constant of integration. If an initial point (x_0, y_0) is specified, we can find the value of the constant.

Example 2

Find the unique solution of the differential equation:

$$\frac{dy}{dx} = 2x + 3\sqrt{x}$$

whose graph passes through $(4, 26)$.

Solution

We have $y = \int 2x + 3\sqrt{x}\,dx$, so that

$$y = x^2 + 2x^{3/2} + C$$

where C is an arbitrary constant. To find the solution through $(4, 26)$, let $x = 4$ and $y = 26$ in the last equation:

$$26 = 4^2 + 2\,4^{3/2} + C = 16 + 16 + C = 32 + C$$

so that $C = 26 - 32$. Thus the function we seek is $y = f(x) = x^2 + 2x^{3/2} - 6$.

An alternative way to obtain Eq. (6) from Eq. (5) is to rewrite Eq. (5) as an equivalent integral equation:

$$\int dy = \int g(x)dx \tag{7}$$

We then integrate the left-hand side of Eq. (7) with respect to y and the right-hand side with respect to x. This yields

$$y + C = \int g(x)dx$$

which is clearly equivalent to Eq. (6).

This particular device leads to the solution for our second type of first-order differential equation.

Case 2 (Variables Separable) Suppose $F(x, y) = g(x)h(y)$—that is, F can be written as the product of a function of x and a function of y. In this case, we say *the variables separate*.

When the variables separate, the differential equation $\frac{dy}{dx} = F(x, y)$ can be written as

$$\frac{dy}{dx} = g(x)h(y) \tag{8}$$

To solve the equation, we write down the corresponding integral equation

$$\int \frac{1}{h(y)}\,dy = \int g(x)\,dx \tag{9}$$

and carry out the indicated integrations.

EXAMPLE 3

Solve the differential equation of Example 1:

$$\frac{dy}{dx} = 3x^2y + \frac{y}{2\sqrt{x}} = y\left(3x^2 + \frac{1}{2\sqrt{x}}\right) \tag{2}$$

Solution

We write the integral equation

$$\int \frac{1}{y}dy = \int 3x^2 + \frac{1}{2\sqrt{x}}dx$$

The next step is to carry out the integrations indicated:

$$\log y + C_1 = x^3 + \sqrt{x} + C_2$$

Recall that log refers to the natural logarithm. A better form for this equation is

$$\log y = x^3 + \sqrt{x} + C$$

Note that we have not yet found y as an explicit function of x, but we have "solved" the differential equation to the extent that we have found a relationship between y and x, which contains no derivatives. In this particular case, we may go further by exponentiating each side of the last equation. We obtain

$$e^{\log y} = y = e^{(x^3+\sqrt{x}+C_3)} = e^{(x^3+\sqrt{x})}e^{C_3}$$

or, more simply,

$$y = Ce^{(x^3+\sqrt{x})}$$

In carrying out the details of this solution, we have assumed that y is strictly positive. A similar result is obtained if y is negative. What happens if y is zero?

Case 3 (The Linear Equation) Suppose $F(x, y) = q(x) - yp(x)$ where q and p are continuous functions of x. Since y occurs only to the first power, this is called a *linear* differential equation. The expression for $F(x, y)$ is linear in y, though not necessarily linear in x.

EXAMPLE 4

Consider the differential equation:

$$\frac{dy}{dx} + 2xy = 2x\sin x + \cos x \tag{10}$$

This equation is rewritten as

$$\frac{dy}{dx} = (2x\sin x + \cos x) - 2xy$$

from which we see that it is a linear differential equation with

$$q(x) = 2x\sin x + \cos x \text{ and } p(x) = 2x$$

I will now show you how to solve this particular linear equation before tackling the general case. What we do may seem strange and terribly unmotivated, but it has the advantage of working. (Motivation can be given, but we want this section to be relatively short.)

Multiply each side of Eq. (10) by e^{x^2}. Since $e^{r(x)}$ is positive for all values of x for any function $r(x)$, we obtain an equivalent equation:

$$e^{x^2}\frac{dy}{dx} + 2xe^{x^2}y = 2xe^{x^2}\sin x + e^{x^2}\cos x$$

At first appearance, we have not improved the situation. It's the second look that counts: the left-hand side of Eq. (11) is precisely a derivative—in fact,

$$e^{x^2}\frac{dy}{dx} + 2xe^{x^2}y = \frac{d}{dx}(e^{x^2}y)$$

and so Eq. (11) may be rewritten as

$$\frac{d}{dx}(e^{x^2}y) = 2xe^{x^2} + e^{x^2}\cos x \tag{12}$$

Now we may integrate each side of Eq. (12) with respect to x. The result is

$$e^{x^2}y = \int (2xe^{x^2}\sin x + e^{x^2}\cos x)dx \tag{13}$$

or

$$e^{x^2}y = e^{x^2}\sin x + C$$

Simplifying, we obtain

$$y = \sin x + Ce^{-x^2} \tag{14}$$

If we were asked to find the solution of Eq. (10) passing through $(0, -2)$, we would simply set $x = 0$ and $y = -2$ in Eq. (14) to compute C:

$$-2 = 0 + C = C$$

The unique solution of the differential equation passing through $(0, -2)$ is

$$y = \sin x - 2e^{-x^2}$$

We are now ready to handle the general first-order linear differential equation:

$$\frac{dy}{dx} + p(x)y = q(x) \tag{15}$$

The first step is to multiply each side of Eq. (15) by the *integrating factor* $e^{\int p(x)dx}$. Then the left-hand side of the resulting equation is an exact derivative and the differential equation can be written in the form

$$\frac{d}{dx}(e^{\int p(x)dx}y) = e^{\int p(x)dx}q(x) \tag{16}$$

Integration with respect to x yields the solution

$$e^{\int p(x)dx}y = \int e^{\int p(x)dx}q(x)dx \tag{17}$$

It will be helpful to examine one final example of a linear differential equation.

EXAMPLE 5

Find the solution of the differential equation $\frac{dy}{dx} = x + y$ passing through $(0, 4)$.

Solution

Rewrite the differential equation in the form

$$\frac{dy}{dx} - 1y = x$$

from which we recognize that $p(x) = -1$ and $q(x) = x$. The integrating factor is $e^{\int -1 dx} = e^{-x}$. Multiplication of the rewritten differential equation by e^{-x} gives

$$e^{-x}\frac{dy}{dx} - e^{-x}y = xe^{-x}$$

which may be reorganized as

$$\frac{d}{dx}(e^{-x}y) = xe^{-x}$$

Integration of each side with respect to x gives

$$e^{-x}y = -e^{-x}(1+x) + C$$

so that

$$y = Ce^{x} - (1+x)$$

Since we are given that $y = 4$ when $x = 0$, we have $4 = C - (1 + 0)$; thus, $C = 5$. Hence, the unique solution of $dy/dx = x + y$ through $(0, 4)$ is

$$y = 5e^{x} - (1+x)$$

Implicit Solutions

In all the examples presented here, the solution techniques described led to an explicit formula for y in terms of x. This is not always possible and, even when possible, is not necessarily desirable. Consider, for example, the differential equation $dy/dx = -2x/y$. This is an example of an equation in which the variables separate. Integration and simple rearrangement yields

$$y^2 = C - 2x^2$$

The solution to the original differential equation is either $y = \sqrt{C - 2x^2}$ or $y = -\sqrt{C - 2x^2}$ depending on the sign of the second coordinate of the initial point. In this example, it is more useful to consider the implicit relation between x and y:

$$2x^2 + y^2 = C$$

from which we see immediately that the points on the solution curve lie on an ellipse centered about the origin.

Other Differential Equations

We have presented solution techniques for only three particular types of first-order differential equations of the form $dy/dx = F(x, y)$. There are other large classes of functions $F(x, y)$ for which exact solutions can be obtained; some of these are discussed in the books listed in the References.

The type of differential equation we have discussed is called a *first-degree* equation because the derivative dy/dx appears only to the first power. Higher-degree differential equations can also be studied. The equation

$$\left(\frac{dy}{dx}\right)^3 + \sin x\left(\frac{dy}{dx}\right) - \log x = e^{-x^2}$$

is an example of a third-degree equation.

Another way differential equations are classified is according to the highest order of differentiation that occurs. For example, the differential equation

$$\cos x\frac{d^2y}{dx^2} + \tan^{-1}x\frac{dy}{dx} + \frac{1}{1+x^2} = 0$$

is a second-order differential equation.

Three of the major areas studied in differential equations are:

1. Techniques for solving various special types of equations
2. Approximation methods to obtain numerical solutions to equations that cannot otherwise be solved
3. Theoretical results on the existence, uniqueness, and qualitative behavior of solutions

This appendix deals with the first topic, and an example of the second topic is given in Chapter 2, whereas Chapter 4 illustrates what can be done in the third area.

EXERCISES

Find the unique solution of each of the following differential equations of the form $dy/dx = F(x, y)$ passing through the point (x_0, y_0). The function F is stated first, followed by the initial point.

1. $x; (2, -3)$
2. $\cos x; (0, 1)$
3. $1/(1+x^2); (1, 0)$
4. $2y; (0, 2)$
5. $2xy; (0, 2)$
6. $x/y^2; (1, 0)$
7. $y/x; (1, 1)$
8. $xy + x; (0, 0)$
9. $x^2 - (y/x); (1, 0)$
10. $\sin x - y; (0, 2)$

You can find a listing of references and suggestions for additional reading on the book's website, www.wiley.com/college/olinick

Index