EDITED BY

Andrew Gelman

AND

Jeronimo Cortina

A QUANTITATIVE TOUR OF THE SOCIAL SCIENCES

CAMBRIDGE

CAMBRIDGE www.cambridge.org/9780521861984

This page intentionally left blank

A QUANTITATIVE TOUR OF THE SOCIAL SCIENCES

Social scientists become experts in their own disciplines but aren't always familiar with what is going on in neighboring fields. To foster a deeper understanding of the interconnection of the social sciences, economists should know where historical data come from, sociologists should know how to think like economists, political scientists would benefit from understanding how models are tested in psychology, historians should learn how political processes are studied, psychologists should understand sociological theories, and so forth.

This overview by prominent social scientists gives an accessible, nontechnical sense of how quantitative research is done in different areas. Readers will find out about models and ways of thinking in economics, history, sociology, political science, and psychology, which in turn they can bring back to their own work.

A QUANTITATIVE TOUR OF THE SOCIAL SCIENCES

Edited by

ANDREW GELMAN

Columbia University

JERONIMO CORTINA

University of Houston

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore, São Paulo

Cambridge University Press
The Edinburgh Building, Cambridge CB2 8RU, UK
Published in the United States of America by Cambridge University Press, New York

www.cambridge.org
Information on this title: www.cambridge.org/9780521861984

First published in print format 2007

ISBN-13 978-0-511-53068-5 eBook (Adobe Reader)

ISBN-13 978-0-521-86198-4 hardback

ISBN-13 978-0-521-68003-5 paperback

Para Zacky, Jacobito, Jacinto, y Camilo

Contents

Preface: Learning to Think Like a Social Scientist

MOTIVATION FOR WRITING THIS BOOK

This book has been created as a one-stop introduction to quantitative empirical social science for busy students. It originated as lecture notes from a one-semester class taught by five professors, with three weeks each of history, economics, sociology, political science, and psychology. Rather than emphasizing common features of all social sciences, we pull out examples that demonstrate the distinctive ways of working within each area. What does it mean to think like an economist? Or like a sociologist, political scientist, psychologist, or historian? We hope that students, in reading this book and working through its examples, will develop some social science literacy in these different ways of thinking, which in turn should improve their understanding of their individual areas of study.

We focus on quantitative models and methods for two reasons. First, quantitative skills are difficult and are in demand, both within and outside academia. Hence, students should be motivated to learn the relevance of quantitative ideas in the study of society. Second, by centering on quantitative approaches, we give a common theme to the book so that students can see how a core group of ideas is applied in different ways to different problems.

MATHEMATICAL LEVEL

This book is not limited to students with a statistical or methodological focus; we think that all social science students would benefit from it, and we are careful to place technical terms in social science contexts.

A unique feature of the book is that we work with quantitative models and methods, but the material is presented in an almost entirely nontechnical fashion. In contrast to books on research methods, ours is a book on *social science*, and we consider the quantitative methods as tools for understanding social phenomena. By reading the book, a student will develop an understanding of the ways in which working social scientists use quantitative models and methods.

The statistical and mathematical ideas considered in different parts of the book are presented intentionally at different levels of complexity. The history chapters discuss the sources of quantitative historical and economic data and the diverse

efforts required to collect and understand this information. In contrast, the chapters on economics focus in depth on a particular area – time series econometrics – to illustrate general issues of evaluating social science hypotheses using quantitative data. The next part of the book, on sociology, begins with a general introduction to social science modeling and then continues with several examples of research indicating the use of statistical models – including how research proceeds when a hypothesized model does *not* fit the data. The political science part has a similar structure, and as with the earlier parts, there is extensive discussion of the practicalities of collecting and evaluating historical data. The psychology part presents a number of theories from social and cognitive psychology along with a discussion of the possibilities of evaluating these theories using empirical data. The book concludes with a discussion and an example of causal inference in social science research.

USING THIS BOOK IN A COURSE OR FOR INDEPENDENT STUDY

We expect that the primary audience for this book will be undergraduates in social science survey courses, junior and senior social science majors, and beginning graduate students. All these students should be interested in an overview of the social sciences with a focus on research methods, whether to learn about methods that they can use in their own senior theses or research projects or simply to get a better understanding of how quantitative knowledge is achieved in these areas.

We envision the use of the book in three sorts of classes. First, it can be a main text (or one of a small set of texts) in a general social science course. Second, it can serve as a supplementary text in a course on quantitative methods in any of the social sciences. For example, in a quantitative methods course in sociology, economics, or political science, it would give students a broad view of the different social sciences in a serious, scholarly, yet easy-to-read format. An instructor can assign this book, along with some homework assignments and focused reading, to give the students a taste of the different quantitative perspectives in the social sciences. The third audience we see for this book consists of students taking social science overview courses in professional programs such as law, public policy, and business.

ACKNOWLEDGMENTS

This project originated from a course that we designed specially for the Quantitative Methods in Social Sciences program at Columbia University. The professors were experienced teachers who enjoyed having an opportunity to give an overview of quantitative methods within their field for a general audience. We thank everyone at Columbia who has helped design and administer the QMSS program, including (in alphabetical order) Peter Bearman, Anthony Cruz, Daphne Estwick, David Krantz, Steven Laymon, Gillian Lindt, Eduardo Macagno, Henry Pinkham, Seymour Spilerman, Tanya Summers, Beatrice Terrien-Somerville, Francis Tuerlinckx, Cynthia Van Ginkel, Greg Wawro, Chris Weiss, and Danielle Wolan. We thank Jorge Balan of the Ford Foundation for arranging funding for

supporting the QMSS program in various ways, including helping to support the costs of transcribing these lectures. We also thank Hayward Alker and Jennifer Hill for helpful discussions, Chandler Sims and Amalia Mena for helping to put this book together, Charles Stockley for his immense effort in transcribing dozens of hours of lectures, and the National Science Foundation and the Columbia University Applied Statistics Center for financial support. Lauren Cowles and her colleagues at Cambridge University Press have provided crucial motivation and assistance in putting together the final version of this book. We are eternally grateful to our spouses, Carolina and Daniela, for their love and support. Finally, we thank the past, current, and future students who have motivated this work and continue to motivate us to combine insights from the different social sciences to study the problems of people in social settings.

About the Contributors

Charles Cameron is a professor of political science at Princeton University. He has been a Fellow at the Brookings Institution, and his book *Veto Bargaining* won the Fenno Prize from the American Political Science Association. He works on applied formal theory and empirical tests of game-theoretic models of politics as applied to legislatures, courts, and elections.

Richard Clarida is a professor of economics at Columbia University and has published more than 25 articles on topics ranging from exchange rates to public investment. He has served on the staff of the President's Council of Economic Advisors and as the U.S. assistant secretary of the treasury for economic policy.

Jeronimo Cortina is an assistant professor of political science at the University of Houston. He specializes in political behavior, immigration, and quantitative methods. His work has been published in scholarly and policy journals such as the *American Politics Research Journal, Foreign Affairs in Spanish*, and the *Harvard Journal of Hispanic Policy*.

Andrew Gelman is a professor of statistics and political science, director of the Applied Statistics Center, and founding director of the Quantitative Methods in Social Sciences program at Columbia University. His books include *Bayesian Data Analysis, Teaching Statistics: A Bag of Tricks,* and *Data Analysis using Regression and Multilevel/Hierarchical Models*. He has also published more than 200 research articles in statistical theory, methods, computation, and applications in various topics in social and environmental sciences.

Emanuele Gerratana received his Ph.D. thesis in economics at Columbia University on the topic of competing mechanisms and common agency. He has also written on family effects of youth unemployment in Italy. He is currently on the faculty of economics at Koc University in Istanbul.

Heidi Grant completed her Ph.D. in psychology at Columbia University and is on the faculty of psychology at Lehigh University, doing research in social psychology and motivation.

E. Tory Higgins is the Stanley Schacter Professor of Psychology at Columbia University and has received several awards for his research in social psychology, including the Thomas M. Ostrom Award for outstanding contributions in social cognition, the William James Fellow Award for Distinguished Achievements in Psychological Science from the American Psychological Society, and the American Psychological Association Award for Distinguished Scientific Contributions.

Herbert S. Klein is a professor of history at Stanford University. He has written 15 books, including *The Atlantic Slave Trade*, *Bolivia: The Evolution of a Multi-Ethnic Society* and *Transicao Incompleta: Brasil desde 1945*, along with more than 125 articles in six languages on topics including comparative social and economic history, international migrations, and demographics of the late colonial to modern periods.

Marta Noguer completed her Ph.D. in economics at Columbia University on the topic of capital flows, technology diffusion, and economic growth. She currently is at the Research Department of la Caixa in Barcelona.

Seymour Spilerman is the Julian C. Levi Professor of Social Sciences and director of the Center for the Study of Wealth and Inequality at Columbia University. He is a Fellow of the American Association for the Advancement of Science and has published more than 70 articles on diverse topics such as occupational attainment, intergenerational financial linkages, racial disorders, and mathematical sociology.

Charles Stockley received his Ph.D. in history at Columbia University on the topic of British economic policy in the 1960s and was previously assistant vice president at Morgan Bank and a vice president at Commodities Corporation.

Elke Weber is a professor of psychology and management and director of the Center for Decision Sciences at Columbia University. She has served as president of the Society for Mathematical Psychology and the Society for Judgment and Decision Making and as a member of a National Academy of Sciences committee on the human dimensions of global change. Her research topics include cross-individual and cultural differences in risk taking.

PART I. MODELS AND METHODS IN THE SOCIAL SCIENCES

Andrew Gelman

1. Introduction and Overview

WHAT THIS BOOK IS ABOUT

The seven parts of this book cover different examples of quantitative reasoning in social science applications, along with interspersed discussions on the role of quantitative theories and methods in understanding the social world. We are trying to show how researchers in different social sciences think about and use quantitative ideas. These approaches vary across the different disciplines, and we think that the best way for a student to get a broad overview is to see how quantitative ideas are applied in a variety of settings.

Hence, we focus on applications, including combat in the First World War, demographics in nineteenth-century Mexico, forecasts of economic conditions by the Federal Reserve, racial disturbances in the 1960s, voting in congressional committees, Pavlovian conditioning, and the effect of migration on solidarity. Each part of the book has several examples, with discussion of the theories used to explain each phenomenon under study, along with the data and models used to describe what is happening. We thus hope to give a general perspective of how these models and methods are used in the social sciences. It is not the usual straightforward story of the form, "Here are some data, here's the model, and look how well the method works." Rather, we explore the strengths but also the limitations of the models and methods in the context of the real problems being studied. Mathematical models and statistical methods can be powerful – you can fit a linear regression model to just about anything – but one needs some sense of what the models mean in context.

The style of the book is conversational and episodic. We do not intend to cover every mathematical model and every statistical method, but rather to give a specific sense of the diversity of methods being used. We believe that the reader of this book or the student in a course based on this book will gain a broader view and some specific tools that he or she can use in attacking applied problems with a social science dimension. This can be especially valuable for a student trained in a specific social science who has only a vague sense of what ideas are used in other fields. A psychological view can be important in understanding a problem that was thought of as pure economics, political scientists need to appreciate the strengths and weaknesses of their sources of historical data, social psychologists

need to comprehend the large-scale studies of sociologists, economists need to think seriously about the workings of governments, and so forth.

Just to be clear: This is not a statistics or a research methods book but rather a social science book with a quantitative focus. We certainly think that social science students need to learn some statistics, and we see the book's ideas as complementing that study. We do discuss linear regression (in Part III) and Poisson models (in Part IV), but we focus on how these answer (or fail to answer) the substantive questions in the particular applications considered. We also discuss some purely mathematical models (for example, game theory in Chapter 3 and decision theory in Chapter 20).

GOOD SOCIAL SCIENCE IS INTERDISCIPLINARY

We don't attempt to define social science except to say that whatever it is, it is studied from different perspectives by historians, economists, sociologists, political scientists, and psychologists. Everybody has his or her own turf, and that's fine, but as a student – or, for that matter, as a practitioner – it's good to know your way around the farm. To some extent, this already happens – quantitative political scientists learn economics, economists use results from cognitive psychology, and just about everybody should learn some history. In this book, we're trying to lay out some concepts in each of the social sciences that can be part of everybody's toolkit in understanding the social world.

We focus on quantitative methods because that's a common ground, an objective reality on which we can all agree – or, where we disagree, we can focus our disagreements on specific questions that, at least in theory, can be answered by hard data. This point arises in Part II of the book; for example, ideas about families that many of us take for granted are associated with demographic patterns in Europe hundreds of years ago. In Part III, there's a discussion of some methods used to assess how much information different economic actors have in predicting exchange rates, and Part IV describes a model for promotion of employees within a large corporation. Different methods can be used to attack similar problems. In political science there's a lot of research on the topic of bargaining, which can be studied quantitatively by looking, for example, at how legislators vote on different issues and to what extent these votes are consistent with political ideology. Part VI discusses several different kinds of theories used to explain decisions and behavior. Some of these theories are more quantitative than others, and it is important to understand under what circumstances such theories can be tested. This issue arises in all the social sciences.

Why do the social sciences use such different methods? One reason is that people study different phenomena in different fields; for example, a discrete Markov transition model[1] might be reasonable to describe changes in a person's

[1] In a "discrete Markov model," a person can be in one of several *states* (for example, 1, 2, or 3, corresponding to "unemployed," "underemployed," or "employed"). The person starts in a particular state, and then at each of a sequence of time points (for example, every month), he or she can stay in the same state or switch to a different state, with specified "transition probabilities" corresponding to each pair of states. Examples of Markov models appear in Parts IV and V of this book, and they can

employment status but inappropriate for modeling gradual changes in public opinion. Another reason for the differences might be personal tastes, traditions, and even political ideology. In any case, if you can understand and communicate in these different fields, you can move to integrate them in solving applied problems, which are never confined to a single academic discipline. Part VII, the last part of this book, provides a more comprehensive, although by no means exhaustive, discussion of causal inference in the social sciences.

CONTROVERSIES IN ESTIMATING THE DOLLAR VALUE OF A LIFE

Mathematical and statistical models can be extremely effective in studying social situations, especially when many methods are used and none is relied on entirely. To take just one case, the economist Peter Dorman wrote a book in 1996 on workers' compensation and valuing human life. The value of a life has long been a contentious topic. From one perspective, it seems immoral to put a dollar value on lives, but in practice this must be done all the time, in settings ranging from life insurance to building code regulations to the cost of air bags in automobiles. For workers' compensation (insurance payments for on-the-job injuries), it is sometimes suggested that dollar values be set based on workers' own valuations of job risks, as revealed by the "risk premium" – the additional amount of money a person will demand in order to take on a risky job. This risk premium can itself be estimated by statistical analysis – for example, running a regression of salaries of jobs, with the level of risk as a predictor variable. In his book Dorman discusses the history of economic analyses of this problem, along with problems with the current state of the art. (For one thing, jobs with higher risks pay *lower* salaries, so the simplest analysis of risks and salaries will "show" that workers actually prefer riskier jobs. An appropriate analysis must control for the kind of job, a perhaps impossible task in practice.)

But what really makes Dorman's book interesting is that he includes historical, political, and psychological perspectives. Historically, workers' compensation has been a government solution to a political struggle between unions and management, so nominal dollar values for risks have been set by political as well as economic processes. In addition, psychologists have long known that attitudes toward risk depend strongly on the sense of responsibility and control over the dangers, and it should be no surprise that people are much less willing to accept new risks than to tolerate existing hazards. Dorman puts all this together, along with some theoretical economic analysis, to suggest ways in which labor,

apply to entities other than persons (for example, a plot of land could be unused or used for residential, agricultural, industrial, commercial, or other purposes, and it can move between these states over time). Such models can be relevant for studying the progress of individual persons or chains of events, but they are generally less appropriate for studying gradual behavior of aggregates. For example, one might consider modeling public opinion in the United States as supporting the Democrats or the Republicans or as balanced between the parties. Such a characterization would probably not be as useful as a time-series model of the continuous proportion of support for each of the parties.

management, and government could negotiate reasonable compromises in assignment of risks.

Dorman's particular interdisciplinary approach is controversial, however. In many different studies, economists have attempted to refine the regressions of wages on risk (these attempts are considered an example of "hedonic regression," so called because they estimate individuals' preferences or judgments about what will make them happiest) to adjust for differences among the sorts of people who take different jobs (see, for example, Costa and Kahn 2004; Viscusi 2000).

As Dorman points out in his book, these questions are politically loaded, and he and others are necessarily giving their own perspectives. But I still believe that the quantitative analyses are useful as long as we are able to evaluate their underlying assumptions – both mathematical and historical.

FORECASTING AS A SOCIAL SCIENCE PROBLEM AS WELL AS A STATISTICAL PROBLEM

I'd like to try to separate the details of a statistical forecasting model from what we learn simply from the existence of a successful forecast. This general idea – that quantifying your uncertainty can reveal the underlying structure of a system – is important in all the observational social sciences, most notably economics, and also is used in engineering in the area of quality control.

Forecasting is an important – in some ways fundamental – task in economics, as discussed in Part III. In addition to its importance for practical reasons (for example, to estimate costs and set prices), the limits of your ability to forecast tell you how much information you have about the underlying system that you are studying. More precisely, the predictive error of your forecast reflects some combination of your ignorance and inherently unpredictable factors (which will always occur as long as humans are involved). Economists sometimes talk about an ideal forecast, which is not a perfect, zero-error prediction but rather is the most accurate prediction possible given the inherent variation in a system between the current time T_0 and the time T_1 about which you are trying to forecast.

To put it another way – this may be a familiar idea if you've studied linear regression – the residual error or unexplained variance represents the variation in the outcome that is not explained by the predictors. For example, a regression of income on job category, years of experience, and an interaction between those predictors will not be perfect – that is, its R-squared will be less than 100% – which makes sense since other factors affect your income (in most jobs). From a social science perspective, the very fact that the regression predictions are accurate only to a certain level is informative, because it tells us how much other factors affect income in this particular society. Statistically, this is one of the ideas motivating the analysis of variance, which involves studying how important different factors are in explaining the variation we see in the world. In addition, the coefficients in a regression model can tell a story by revealing which predictors carry information, which is discussed in detail at the end of Chapter 9.

To return to time-series forecasting, if a variable can be accurately predicted two years ahead of time, then intervening factors during those two years cannot be important in the sense of affecting the outcome. What is more common is that a

forecast several years in advance will have a large error variance, but as the time lag of the forecast decreases, it will become increasingly accurate – but with some limit on the accuracy, so that a very short term forecast (for example, of tomorrow's prices) will still not be perfect.

From a social science point of view, the situations and time lags where a forecast has large errors provide an opportunity to try to understand the intervening variables that make a difference. (A similar point comes up in Part IV, where data on racial disturbances do *not* follow a Poisson model, motivating a search for further understanding.) We consider this in more detail for election forecasting.

There are many ways to predict the outcomes of elections. In the following discussion we focus on U.S. congressional and presidential elections, but the same ideas apply to other contests in the United States and other places where elections are held regularly. You can forecast two years ahead of time, or three months before the election, or one week before, or on election night. Each of these forecasts can be useful for different purposes, and each uses different information. For example, to forecast an election two years early, you'll use past election results, forecasts of the economy in two years, and general political knowledge such as that the president's party often does poorly in off-year elections. Three months before the election the candidates are known, so you can include information on candidate quality. You also have some information from polls, such as presidential approval ratings and answers to the question "Who would you vote for?" as well as more reliable economic data. These forecasts have received a lot of attention in the economics and political science literature (studies by Ray Fair [1978], Steven Rosenstone [1983], Bob Erikson and Chris Wlezien [1996], James Campbell [1996], Michael Lewis-Beck and Tom Rice [1992], and others) because of their inherent interest and also because of the following paradox: The final months of election campaigns sometimes feature dramatic reversals, but if elections can be accurately forecast three months ahead of time, this seems to allow little room for the unexpected. As Gelman and King (1993) put it, why do preelection polls vary so much when elections are so predictable? In this setting, the very existence of a forecasting model has political implications.

To continue briefly with the election forecasting example, the next step is prediction on election night itself. Partial information – the election results from the first precincts and states to report – can be used to forecast the whole. Time-series forecasts are used in an interesting way here: From a previously fitted model, you can get a forecast for each precinct and state. As actual election returns come in, you can see where each candidate is performing better or worse than predicted and then make inferences about the forecast error – and, from there, the actual election results – in the rest of the country. The important fact, from a modeling standpoint, is that these cross-sectional errors are in fact correlated and that correlation is part of the model used to make the forecast.

Although interesting as a statistical problem and important to news organizations, election night forecasting is not particularly interesting from a political or economic perspective. It is certainly no surprise that elections can be forecast from partial information, and, in any case, you can always just wait for the morning to find out who won. An exception was the 2000 election, where

forecasting models had a sort of afterlife, being used, among other things, to show the implausibility of Patrick Buchanan's vote total in Palm Beach County, Florida (see, for example, Adams 2001).

The example of election forecasting illustrates how similar statistical models can have different social science implications. U.S. presidential elections can be forecast fairly accurately months ahead of time. This implies that campaigns, in aggregate, have little effect and in some ways seems to imply that our vote choices are less the product of free will than we might imagine. Social scientists found similar results decades ago (Converse 1964): From a series of opinion polls they found that voters' preferences were very stable, even across generations; people inherit their political views much as they inherit their religion (or, for that matter, their height and body type). In contrast, the accuracy of election night forecasts, based on a decomposition of vote swings at the national, regional, state, and local levels, is somewhat reassuring in that these forecasts reinforce the idea that the United States is a single political entity but with local characteristics. Once again, similar results have been found in public opinion polls: Page and Shapiro (1992) use the term "parallel publics" to refer to the phenomenon that national changes in opinion tend to occur in parallel throughout different segments of the population. Similar models are used for tracking information in other social sciences but to different ends: for example, the study of individual judgment and decision making in psychology and the study of information flow and communication in economics.

OVERVIEW OF THE BOOK

This book is about mathematical models; ways of understanding quantitative data; concepts of probability, uncertainty, and variation; and sources of historical and current data – all applied to the understanding and solution of social problems. With parts on each of five social sciences – history, economics, sociology, political science, and psychology – we teach by example some of the different ways in which researchers analyze social phenomena using mathematical ideas and quantitative data.

Although this is not a statistics textbook, we refer throughout to the use of statistical methods. Our focus is on the social problems being modeled. Also, we are not attempting to present a unified quantitative view of the social sciences; rather, we favor a more pluralistic approach in which different methods are useful for different problems. This book is intended to make the reader comfortable with the kinds of quantitative thinking used in the different social sciences.

In the next two chapters, I'll present some specific examples in which quantitative modeling or analysis has had an impact. My focus will be on how the mathematical or statistical methods link to the underlying social science. The rest of the book, as noted previously, is divided into separate parts for each of the major social sciences. The parts have little overlap, which is intentional: We want you to see the different perspectives of the different social sciences.

We continue our tour of the social sciences with history. Much of our understanding of history (including the recent and immediate past) is based on numbers, ranging from national statistics on populations, trade balances, public

health records, and voting records to data that are important for particular studies such as those on dietary consumption, television ratings, and sports attendance. (In between are data sources such as national opinion polls, which are privately collected but used for public policymaking.) In addition, nonnumerical data can be summarized quantitatively; for example, in the study of a historical debate, contemporary newspaper articles for and against the issue can be counted. Part II of this book begins with a discussion of the historical uses of statistics and then considers various primary sources of quantitative historical data. The lecture material is best supplemented with two kinds of exercises: first, those in which students must collate some historical data (for example, from city property records or old newspapers) and, second, those in which they must read secondary sources and summarize a scholarly debate about some quantitative question with historical implications (for example, the number of people living in the Americas in 1492).

We then move to economics, which is increasingly important in a variety of settings, from private-sector decision making (for example, deciding how seriously a consumer products company should take its sales forecasts) to cost-benefit analyses such as evaluating how much the population is willing to pay for public parkland. Part III begins with a general discussion of the difficulties of understanding and using economic data and then presents three examples of analysis of time series of interest and exchange rates. The focus is not on the particular statistical models being used, but rather on the way in which economic concepts are encoded into the models. All the analysis is done using simple least squares or two-stage least squares, and the key concerns are the predictor variables in the equations and the interpretation of the coefficients. One thing all these examples have in common is that they use macroeconomic data to study the flow of information, which is a major theme in modern economics. Doing the exercises for Part III is a good way to learn data analysis on the computer, starting with simple downloading and plotting of time-series data and then fitting some simple time series and regression models. For the model-fitting exercises, it is important to interpret the coefficients in the model in terms of the motivating economic questions in order to have a good understanding of the theoretical implications of the model

Part IV covers sociology, which in many ways is the most flexible and interdisciplinary of the social sciences. At the theoretical level, sociologists consider a variety of models for social interactions and the roles of individuals in groups, with no single underlying set of principles (unlike in economics). In addition, sociologists often invent new methods of analysis to study unusually structured data (for example, on social networks). Part IV begins with a general discussion of social science theories and the ways in which a single phenomenon can be understood from different theoretical perspectives. It then presents two examples of models for discrete events in which quantitative data are used to understand violent incidents in race relations. For these examples, we discuss the challenges of modeling as well as accurately summarizing historical data. Finally, there is a discussion of models for promotion of employees within large organizations. This sort of analysis can be useful in a wide variety of contexts when studying the fates of individuals within a large social structure. The mathematical ideas used in sociology, such as Markov chains and the Poisson distribution, are

not extremely complex, but they are new to many students (who are more likely to be trained in linear regression models). Hence, the exercises for the chapters in Part IV are focused on these models. For example, in one assignment you must come up with a set of count data to which you must fit the Poisson distribution and then comment on any lack of fit of the model.

We continue with Part V on political science, which encompasses both the particular study of individual governmental processes and the general study of political behavior. At a practical level, anyone trying to solve a problem – inside or outside of government – should be aware of the presence of governments at local, state, and national levels, whether as means for solving the problem or as arenas for competing possibilities. More generally, political science – the study of ways to work around conflict – is also relevant in nongovernmental settings. Along with economics, political science as a field can be highly quantitative due to the availability of vast amounts of numerical data (particularly in more industrialized countries) on taxes and spending, voting, and opinion polls. At the same time, political theory can be highly mathematical. Part V contains detailed discussions about how social theories can be constructed so that they can be testable using empirical data. This is followed by specific examples of linear regressions and more sophisticated methods using game theory and Markov chains to model political decision making and voting in committees.

Part VI focuses on psychology, which is both a cognitive and a social science – that is, it studies people's internal processes as well as their interactions with others. Psychology is clearly necessary, at some level, for understanding all social phenomena, since they all are ultimately based on human actions. For example, the laws of micro and macroeconomics and the observed regularities of politics require some sort of consistent behavior, either of individuals or in the aggregate. Part VI starts with a general discussion of theories of human behavior and how these theories can be tested or disproved. It continues with descriptions of some mathematical models of decision making, which in turn suggest where classical economics might be relevant to describing individual behavior and where it might fail. A useful exercise in Chapter 18 is to compare the predictions that several different psychological theories would each make about an actual experiement. Chapter 20 presents some exercises on decision making and the combination of information.

The book concludes with Part VII, a discussion of causal inference in the social sciences. It focuses on the ever-present issue of causal versus spurious relations and discusses the example of estimating the effects of migration on solidarity in migrant-sending communities.

Our chapters go back and forth between general discussions of social science modeling and applied examples that are designed to allow class discussion. The material can be used as a starting point to explore the variety of mathematical and statistical models used in social science, with the ultimate goal of enriching your understanding of the complexities of the social world.

2. What's in a Number? Definitions of Fairness and Political Representation

A key starting point of quantitative social science is *measurement* – which encompasses direct observations from personal interviews and field observations, large-scale data collection efforts such as surveys and censuses, structured observations in designed experiments, and summary measures such as the Consumer Price Index. All these form the raw material for larger social science studies, and it is easy to get lost in these analyses and forget where the numbers came from and, even more importantly, what they represent.

We illustrate the choices involved in numerical measurements in the context of a subject of general interest – how people are represented in a political system – that can be studied both at a theoretical and an empirical level. I want to make the case that quantitative summaries can be helpful, as long as we are aware of the choices that must be made in summarizing a complex situation by a single set of numbers.

We want our political system to represent the voters and treat them fairly. At the simplest procedural level, this means giving a vote to each citizen and deciding elections based on majority or plurality rule. In practice, however, we are not all represented equally by the government, and as long as there is political disagreement, there will be some dissatisfaction. It would be appealing to have a mathematical definition of the amount of citizen "representation." Unfortunately, different measures of representation can interfere with each other, as we discuss with examples from national elections in the United States.

WHAT IS THE MEANING OF POLITICAL REPRESENTATION?

The United States is a representative democracy, and we vote for people who represent us: congressmembers, the president, state legislators and governors, and local officials. Indirectly, through our elected representatives, we vote for the justices of the Supreme Court and other persons in appointed positions.

What does it mean for us to be represented in this political system? For one thing, everyone's vote counts equally, and as a consequence of two Supreme Court rulings in the 1960s, most legislatures are set up so that the number of

people living in each district is about the same.[1] (Two important exceptions are the Electoral College and especially the U.S. Senate, both of which overrepresent small states.)

The right to vote could be thought of as a "procedural" aspect of democracy, but what about actual outcomes? Representation can be of individuals or of groups, and both of these perspectives are relevant. For example, suppose you lived in a country where 90% of the voters got what they wanted in the sense that their favored candidates for Congress won. So, 90% of the people are happy on election night (at least for the congressional elections). That's real representation, right? Well, maybe not.

In this scenario, you'll have very few, if any, close elections. For example, maybe the Democrats get 95% of the vote in one district; the Republicans get 88% in another, and so forth. If 90% of the voters get what they want, this can happen mathematically only if almost all the districts have landslide elections. Then, all the seats are "safe," so the congressmembers have no reason to fear the voters. If they fear us, they'll respect us; otherwise, they can ignore us, and then we have no influence at all! (I don't mean to be cynical here. I assume that most politicians are trying to do what's best for the country, but if they think the election might be close, then they will probably worry a bit more about what *we* think is best for the country, too.)

Electorally, to get the benefit of two-party competition, you need competition in the individual districts, not just in the country as a whole. Hence, there is a whole theory of "pivotal votes" (see Gelman, Katz, and Bafumi 2002 for an overview).

How close are actual elections? Figure 2.1 shows histograms of the Democratic share of the two-party vote for elections to the U.S. House of Representatives in 1948 and 1988. Each histogram represents 435 districts and, as you can see, many are noncompetitive (that is, far from 0.5), especially in 1988. This has been studied in different ways (see, for example, Gelman and King 1994), and there has been a trend in the United States in the past few decades for legislative elections to be less close. Much but not all of this trend can be attributed to an increase in incumbency advantage – it's worth about 10% of the vote to be a sitting member of Congress (see Cox and Katz 1996, Erikson 1971, and Gelman and King 1990, among others). Politicians are not stupid, and when they see elections that are less close, they are less worried about keeping the voters happy.

An exception was in 1994, when there was a big swing to Newt Gingrich and the Republicans. But, even then, in most of the country the elections were not close. For example, I was living in Berkeley, California, at the time, and there was no way that a Republican would win there. Poll results sometimes find that most Americans are dissatisfied with Congress as a whole, but they like their local representative; this is sometimes taken as a paradox, but it makes sense, given that most congressional elections are not close (see Figure 2.1).

So far, we've considered three senses of representation: having equal votes, being satisfied with electoral outcomes, and having your vote have the potential to make a difference. Other measures are possible that are perhaps more important politically.

[1] In their book *The End of Inequality*, Stephen Ansolabehere and James Snyder (2008) discuss the changes in U.S. politics that have occurred since congressional and state legislative districts were required to be drawn with equal populations.

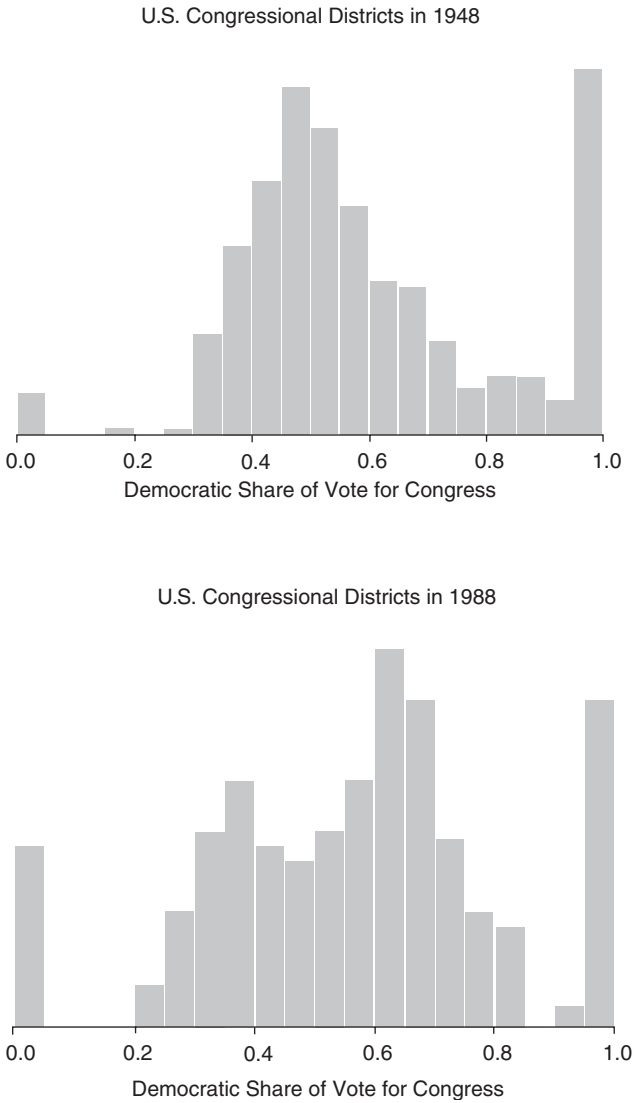

Figure 2.1. Democrats' share of the two-party vote for the 435 elections to the U.S. House of Representatives in 1948 and 1988. (The spikes at the extremes of the graphs represent uncontested elections.) There were relatively few close elections (that is, with vote shares near 0.5) in 1988.

For example, instead of asking if your vote is influential, we could ask if your money is influential. Given the effort put into campaign fund-raising, it is reasonable to suppose that individuals and groups that give more money are better represented by politicians.[2]

[2] Thomas Ferguson, in his 1995 book *Golden Rule*, backs this claim with data on campaign contributions from different industries over several decades of elections.

Table 2.1. *Some comparisons of the U.S. population to the U.S. House of Representatives in 1989 and 2006*

	Proportion in U.S. population		Proportion in House of Representatives	
	1989 (%)	2006 (%)	1989 (%)	2006 (%)
Catholic	28	24	27	30
Methodist	4	7	14	12
Jewish	2	1.3	7	6
Black	12	12	9	10
Hispanic	10	15	4	6
Female	51	51	6	16
Under 25	30	34	0	0

Sources: Adapted from King, Bruce, and Gelman (1995); US Census Bureau http://www.census.gov/popest/estimates.php; http://womenincongress.house.gov/data/wic-by-congress.html?cong=109; www.adherents.com; Congressional Research Service (CRS) Report "Black Members of the United States Congress: 1870–2005" available at http://digital.library.unt.edu/govdocs/crs/permalink/meta-crs-7142; http://www.senate.gov/reference/resources/pdf/RS22007.pdf

A completely different notion of representation is demographic: Do your representatives look like you? Table 2.1 shows some data from the U.S. Congress: Representation varies greatly by religious and ethnic group. The most extreme case, lack of representation of the under-twenty-fives, is a consequence of the minimum age requirement in the Constitution.

A related issue is representation by occupation: Many congressmembers are former lawyers and businesspeople and some are former teachers, but not many are former janitors. My colleague from the Netherlands thought it surprising that in the United States, even the Democrats rarely nominate labor union officials for elective office. He said that in Europe, it is considered unseemly for left-wing parties to run white-collar candidates and that such politicians are called "saloon socialists." I said, that's funny, "saloon" sounds pretty down-to-earth. He replied, oh, that's right, it's "salon" socialists.

Another thing to look at is the representation of your political views. Suppose you support abortion rights, a missile defense system, and a higher minimum wage. If all these policies are being implemented, maybe you should feel happy whether or not the candidates you vote for actually win. One could measure this sort of representation using opinion polls and compare the average satisfaction levels of different groups in the population.

FAIRNESS AND REPRESENTATION

What about the representation of political parties? Most of Europe uses proportional representation: If a party gets, say, 20% of the vote, it is given 20% of the seats in the legislature. In some countries, a party that gets only 2% of the votes still receives 2% of the seats. Other countries have a threshold of 5% before a

party gets representation; they don't want every group with 1% of the vote to have a representative, screaming, putting chewing gum in the elevators, and generally causing problems in the legislature.

The United States does not have proportional representation. There's a separate election in each congressional district, and whoever wins that election goes to Congress. In theory, a party could get 49% of the vote and still get zero representation – if it got 49% in every district in the country. More realistically, a party could get 45% of the vote and only 40% of the seats or 20% of the vote and no seats. Suppose some unfortunate third party gets 20% of the vote in every district. In one district, the Democrat might receive 42% and the Republican 38%. In another, the Democrat might win 60% while the Republican ties with the third party at 20%. Regardless, the third party loses out.

Some people consider the lack of proportionality to be a defect in our system, with the 20% of the votes that went to the third party wasted, since they did not lead to any representation. Of course, you could also consider the 38% for the Republican candidate as wasted or consider all votes for any losing candidate as wasted. For that matter, you can consider extra votes for a winning candidate as wasted, too – did that person really need 60% of the vote? – but it seems particularly tough on that third party, since *all* of its votes are wasted.

But proportional representation has a problem too: Small changes in the vote produce only small changes in seats for the political parties. This is a problem because swings in votes between national elections are typically only about 5%. In the American (or British) system, a 5% swing in votes can easily produce a 10% swing in seats – enough to possibly change which party controls the legislature. This gave Newt Gingrich and the Republicans control of the House of Representatives in 1994. You might not have been happy with that particular outcome, but it's reassuring that a change in votes has the power to change who rules Congress. Under proportional representation this can happen too, but in a more subtle, less voter-controlled way: Perhaps some minor party increases its vote share from 10% to 15%, and then it can make a deal with another party, ultimately changing the government. The change in votes has an input, but not so directly: The proportional representation system with multiple parties is more like a pinball game where the voters shoot the ball, and then the parties keep it bouncing all by themselves.

VOTES

Amid all this discussion of rules, we shouldn't forget the people being represented. A congressional district in the United States has nearly 700,000 people (the number varies slightly from state to state), of whom perhaps 500,000 are eligible voters (over age eighteen, citizens, nonfelons). So, you can win your very own seat in Congress with 250,000 voters, or 100,000 voters if the turnout is 40%. If the turnout is below 20%, then if you can convince the right 50,000 people to vote for you, you can represent all 700,000 people in the district. According to the Constitution, the elected legislator represents everyone who lives in the district, including those who could not vote, did not vote, or voted for the losing candidate.

But obviously, a politician will be less concerned with the people who did not vote for him or her, whatever the reason.

For that matter, my friends in other countries say that they should get to vote for the president of the United States also, since he has such a large effect on their lives.

INHERENT BIASES OF THE POLITICAL SYSTEM

As we have seen, even if our political system is working perfectly as designed, not all individuals and groups will be treated equally. Children don't get to vote and can't hold political office, even though they are nominally represented in the government. Voters in small states are vastly overrepresented in the U.S. Senate. Votes for minor parties are generally wasted (at least in their direct effects), and, as we have seen, proportional representation creates other problems. And as long as campaigns need money, rich people and better-funded groups can expect disproportionate representation of their political views.

Looking at representation in terms of decisive votes creates other paradoxes. If you want politicians to fight for your vote, then elections have to be close (or at least potentially close), but when an important election actually is close (such as the 2000 presidential election), half of the people will feel unrepresented.

Two other systematic biases that have been studied by political scientists are the "tyranny of the majority" and the "median voter" rule. Majority rule has always been considered dangerous since, for example, 51% of the people could get together and vote to tax the other 49% into poverty (which worried conservatives during the New Deal in the 1930s). The founders of the U.S. Constitution created various checks and balances to slow this trend down, but it is still something of a mystery why the majority in a democratic system is not more tyrannical. Perhaps one reason is that most people do not trust politicians enough to lend them this power.

The median voter rule was formulated by Harold Hotelling in 1929 as an application of a theorem in economics and was developed further in a book by Anthony Downs (1957). We illustrate the basic idea in Figure 2.2: The curve represents the distribution of voters in the electorate, ranging from far left to far right, and the usual positions of the Democratic and Republican candidates are indicated by D and R, respectively. Suppose that any voter will choose the candidate who is ideologically closest to him or her. Then all the voters to the left of the D position will go for the Democrat, all the voters to the right of R will side with the Republican, and the voters in the middle will go for whichever is closest.

In this scenario, the Democrat will gain votes by moving to the right – he or she will still get all the voters on the left and will also gain some votes in the center. Similarly, the Republican should move to the left. Ultimately, the only stable position for the candidates is for them both to be at the position of the median voter (labeled M in Figure 2.2). If either candidate deviates from this position, the other can move to the median and get more votes. This explains why presidential candidates tried to sound moderate. It's not that there are no voters at the extreme; it's just that these extreme voters have nowhere else to go. (In 2000, 2004, and 2008, all the minor-party candidates together got less than 5% of the vote.)

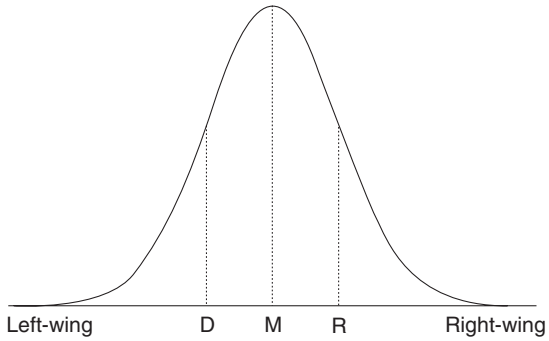

Figure 2.2. Sketch of the theory of voter competition leading to the rule of the median voter. The curve represents the political ideologies of the voters (it is drawn as a bell-shaped curve for convenience but in general can be any distribution). D and R indicate the initial positions of the Democratic and Republican candidates. If all voters turn out and the two candidates have complete ideological flexibility, then they will both end up at the median voter's position, M.

Getting back to representation, this theory suggests that the median voter is strongly represented by the political system, whereas voters away from the center have no representation. This translates into less representation for groups such as African Americans and other ethnic minorities whose political views are far from the center.

In real life, however, Democrats and Republicans do not occupy the same point at the center of the political spectrum, and so the median voter rule does not tell the whole story, as in fact is indicated by the differing initial positions of D and R in Figure 2.2. The differences between parties can be studied in various ways; Figure 2.3 shows some data based on votes in the House of Representatives in 1992. The 435 members of Congress were ranked from left to right based on their roll-call votes (Poole and Rosenthal 1997), and this is plotted versus the liberalness or conservativeness of their districts, as measured by an adjusted version of Bill Clinton's share of the presidential vote in 1992. (The polarization between the two political parties is discussed further in Part V of this book.)

Figure 2.3 shows that Democratic and Republican politicians differ greatly in their ideologies (as measured by their actions in Congress), even after controlling for the political slants of their districts. There are many reasons for this; our point here is that the median voter rule gives some insights into political representation, but it is not completely borne out by the data. Another way to study the influence of median and extreme voters, which we won't get into here, is to look at survey results to see where the positions actually taken by elected politicians fall on the spectrum of public opinion.

DIFFERENT PERSPECTIVES ON REPRESENTATION

Definitions of representation can be categorized in several ways. First, there is the distinction between procedure and outcome. The laws may treat everyone equally, but differences in resources can translate into differences in political power. Second,

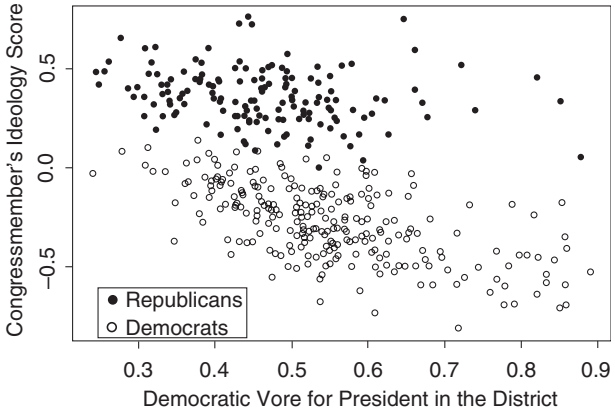

Figure 2.3. Ideology scores (high values are conservative, low values are liberal) for congressmembers in 1992 plotted versus the Democratic proportion of the two-party presidential vote in their districts. (The presidential vote has been corrected for home-state advantage and other effects to yield an estimate of the party's political strength in the district.) Congressmembers in more Democratic-voting districts tend to be more liberal, but there is a dramatic difference between the ideologies of the Democrats. This ideological gap implies that the median voter (see Figure 2.2) is not completely dominant.

one can distinguish between electoral outcomes and ultimate satisfaction with policy. Representation requires political participation at many steps in the political process (see Verba, Schlozman, and Brady 1995). And one can study many levels of government, including local, state, and national, as well as quasi-governmental organizations such as public corporations. We're used to thinking of government as one thing and private enterprise as another, but there's a lot of discussion in the social sciences about how these categories blur. On the one hand, areas of government can act like private businesses – sometimes in a good way, by serving customer demand and allocating resources, and sometimes less beneficially to the public interest, for example by charging the taxpayer for no-bid construction projects. On the other hand, large corporations and labor unions are like governments in that they must balance many different internal interests, and on a personal level they can provide a promise of security that is one of the traditional functions of government. Business, consumer, or special-interest lobbying groups are generally unelected but claim to represent interests in the political process. Internationally, nongovernmental organizations such as the World Bank serve some of the roles of government, but it is not always clear whom they represent.

More fundamentally, representation can be defined in relative or absolute terms. Fairness requires that any person or group be represented just as much as any other, but on an absolute scale it is possible for all the citizens of a country to be more or less represented by its political system. (And, to return to our earlier concerns, it is possible for people to get what they want politically without representation or for well-represented groups to be dissatisfied.)

What's best for you individually might not be best for your group. For example, maybe you would rather not pay taxes, but if nobody pays taxes, we're

all in trouble. This is also true of representation. We'd all like to have more influence over government policy, but at some point it's mathematically impossible – there are more than 300 million of us, and we can't all be decisive. And, as we discussed in the context of proportional representation, a system that seems fair in one way can reduce representation in others. Political theorists have been struggling with these questions for millennia (see Beitz 1990 for a review).

From the standpoint of quantitative social science, it is interesting to see how empirical data, in conjunction with theoretical ideas, can be used to get a better perspective on some issues of representation. Once we go beyond the search for a single numerical measure of fairness, we can study how our institutions represent us, both individually and as groups. When crunching numbers, it's easy to get stuck worrying about relatively minor aspects of the political system, such as gerrymandering and the allocation of congressional seats, and forget about more systematic ways in which people get more or less of a say in their collective institutions.

3. The Allure and Limitations of Mathematical Modeling: Game Theory and Trench Warfare

We conclude the first part of this book with a critical overview of an application of mathematical theory to a problem in history – a game-theoretic model that was developed more than twenty years ago to describe some surprising behavior in the trenches in the First World War. As we shall see, the model is appealing because it seems to bring some sense to a scary and confusing subject. However, ultimately, I find this particular analysis unconvincing, and I'll talk about why such a model, if wrong, was so persuasive. This is an important step in social science analysis, inside or outside the academic setting: If you're going to claim that you're right and other people are wrong, you'd better also explain why those other people (who were generally not fools) got it wrong. Scholarly journals are littered with proudly "counterintuitive" findings that are counterintuitive for the simple reason that they are wrong and most people's intuition is right.

THE PRISONER'S DILEMMA: A GAME-THEORETIC MODEL APPLIED TO SOCIAL INTERACTIONS

Mathematical theory can be a powerful tool for understanding social phenomena but can also mislead, as I'll illustrate with an example from my undergraduate thesis, which was in political science. At the beginning of my senior year in college, I reviewed a list of potential advisors and found one who worked on game theory – I was a physics and mathematics major, minoring in political science, and game theory seemed like a good topic. My advisor handed me *The Evolution of Cooperation*, a then-recent book on game theory, society, and evolution, written by Robert Axelrod (1984). Nowadays, books on evolution and social science – sociobiology – are commonplace, but this book, written by a political scientist and published in 1984, was one of the first and was extremely well received in the general and scholarly press.

I'll describe this example in detail to illustrate how mathematical ideas – in this case, of game theory – can be applied to real social and historical problems. In addition to illustrating the power of mathematical modeling, this example shows some of its weaknesses. (I would summarize Axelrod's book as a valuable theoretical argument with fundamental flaws in its applications.) Finally, we'll explore the political implications of explaining human actions quantitatively. Is it

possible to use mathematics to gain insight without surrendering the political equivalent of free will?

The fundamental problem Axelrod was studying is *cooperation*. From a psychological or economic point of view, why do people cooperate with each other (instead of acting purely selfishly, which would give them short-term benefits, at the very least)? A historian might study cooperation as it has existed in past societies, and a sociologist might consider the settings in which individuals assume cooperative and noncooperative roles. From a political science perspective, the natural question is how to *promote* cooperation – behavior that is essential to the functioning of any political system to avoid a Hobbesian war of all against all. (In the 1640s, Thomas Hobbes wrote *Leviathan*, the classic analysis of political life arising from a brutal state of nature.)

From a game-theoretic standpoint, cooperation has always been viewed as a puzzle, and this puzzle has been given various names. The "free rider problem" refers to the logical motivation of any individual to sit lazily while the heavy lifting is done by others. In a political context, free riding might mean avoiding voting, paying taxes, or military service. The "tragedy of the commons" (Hardin 1968) describes what happens when people behave selfishly – for example, consuming more than their fair share of resources or polluting a common water supply – under the expectation that they might as well act selfishly, since others will do so if they do not. A classic example is overfishing; any given fisher will want to catch as many fish as possible, but if this happens, then all the fish will be caught and everybody loses. Both the free rider problem and the tragedy of the commons have the following properties: (1) if everyone behaves cooperatively, then they will all do well, (2) any individual will do even better by behaving selfishly, but (3) if all (or even many) individuals behave selfishly, all will suffer – even the selfish ones. Social scientists often refer to this as a "prisoner's dilemma game" by analogy to a different cooperation problem that we will not go into here (see the Axelrod book for details if you'd like, or Maurer and Tucker [1983] for the history of the term, which is due to Albert Tucker).

Table 3.1 shows a formal expression of the shared-commons or prisoner's dilemma. The key question here is, why do people cooperate at all in situations like this or, to reframe it more constructively, how can we develop social systems to encourage cooperation? In the long term, cooperation makes sense, but in the short term, it pays to not cooperate. How can cooperative short-term behavior occur? Several answers have been suggested. People are more comfortable cooperating with people they know, and this has been studied experimentally by economists and psychologists (see Dawes, De Kragt, and Orbell 1988). In situations where cooperation is important (for example, in a business) or even a matter of life and death (for example, in the military), it is considered crucial to create a team spirit.

However, in other settings, most notably in the economic sphere (recall the overfishing example), the incentives to not cooperate are so strong that psychological motivation does not seem enough. Cooperation can then be enforced through governmental action or private binding agreements (which themselves typically require governmental presence to be enforceable). Economists refer to these situations where cooperative behavior requires outside enforcement as