Political Science Research Methods

Eighth Edition

Janet Buttolph Johnson
H. T. Reynolds
Jason D. Mycoff
University of Delaware

SAGE | CQPRESS

Los Angeles | London | New Delhi
Singapore | Washington DC

SAGE | CQ PRESS

Los Angeles | London | New Delhi
Singapore | Washington DC

FOR INFORMATION:

CQ Press
An Imprint of SAGE Publications, Inc.
2455 Teller Road
Thousand Oaks, California 91320
E-mail: order@sagepub.com

SAGE Publications Ltd.
1 Oliver's Yard
55 City Road
London EC1Y 1SP
United Kingdom

SAGE Publications India Pvt. Ltd.
B 1/I 1 Mohan Cooperative Industrial Area
Mathura Road, New Delhi 110 044
India

SAGE Publications Asia-Pacific Pte. Ltd.
3 Church Street
#10-04 Samsung Hub
Singapore 049483

Acquisitions Editor: Sarah Calabi
Development Editor: Nancy Matuszak
eLearning Editor: Allison Hughes
Editorial Assistant: Raquel Christie
Production Editor: Olivia Weber-Stenis
Copy Editor: Talia Greenberg
Typesetter: C&M Digitals (P) Ltd.
Proofreader: Theresa Kay
Indexer: Sheila Bodell
Cover Designer: Gail Buschman
Marketing Manager: Amy Whitaker

Printed in Canada

ISBN 978-1-5063-0782-4

This book is printed on acid-free paper.

FSC www.fsc.org MIX Paper from responsible sources FSC® C011825

15 16 17 18 19 10 9 8 7 6 5 4 3 2 1

Brief Contents

Detailed Contents

Tables, Figures, and Features

Tables

Figures

Helpful Hints

How It's Done

Preface

A political science student may ask, "My interest is government and politics; why do I have to study research design, question wording, document analysis, and statistics?" Our goal in *Political Science Research Methods* is to address this question by demonstrating that with a modicum of effort applied toward studying these topics, undergraduates can analyze many seemingly complicated political issues and controversies in ways that go far beyond accounts in the popular press and the political arena.

Political Science Research Methods, now in its eighth edition, continues to hold true to the three primary objectives that have guided us since the book's inception. Our first objective is to illustrate important aspects of the research process and to demonstrate that political scientists can produce worthwhile knowledge about significant political phenomena using the methods we describe in this book. To show this as vividly as possible, we begin again with several case studies of political science research drawn from different areas of the discipline that address key issues and controversies in the study of politics. We made an effort in this edition to include a wide variety of examples from the main subfields of political science. We continue to make changes to fulfill our other two objectives: (1) to give readers the tools necessary to conduct their own empirical research projects and evaluate others' research, and (2) to help students with limited mathematical backgrounds understand the statistical calculations that are part of social science research. Though we are increasingly concentrating on what various procedures can (and cannot) tell us about the real world, we've tried to include examples of procedures and their associated calculations most likely to be used by students. We still provide separate computational details from the narrative by placing many equations in "How It's Done" boxes. The book makes an effort to encourage students to understand and think about the practical and theoretical implications of statistical results. We hope that by meeting these goals, this book will continue to satisfy the needs of our undergraduate and graduate students as they embark on their studies in the field.

Structure and Organization of the Book

In this eighth edition we have responded to feedback that called for a tighter focus on what instructors say matters most. We carefully streamlined each chapter to deliver greater clarity of concepts and added new learning objectives to encourage close reading of main takeaways. In addition, a new, colorful interior visually highlights the content's accessibility.

Because research methods may overwhelm some students at first, we have gone to some length (in the first chapter, especially, but also throughout the book) to stress that research methods topics can be relevant to the understanding of current events. This book is organized to show that research starts with ideas and then follows a series of logical steps. Chapter 1 introduces the case studies that are integrated into our discussion of the research process in the subsequent chapters. We chose these cases, which form the backbone of the book, to demonstrate a wide range of research topics within the discipline of political science: American politics, public administration, international relations, comparative politics, and public policy. We refer to these cases throughout the book to demonstrate the issues, choices, decisions, and obstacles that political scientists typically confront while doing research. We want to show what takes place behind the scenes in the production of research, and the best way to do this is to refer to actual articles. The advantage to this approach, which we feel has been borne out by the book's success over the years, is that it helps students relate substance to methods. For this edition, we added a new example of research into the gender gap in politics, which is especially useful as it demonstrates the use of direct observation as a data collection method. We updated and extended the example on income inequality and redistribution in Organisation for Economic Co-operation and Development (OECD) countries and its causes to include research into the relationship between income inequality and political representation in the United States. This topic still links nicely to the example on voter turnout. The example of control of the bureaucracy remains as before. Our discussion of research into human rights abuse continues to track the evolution of that topic. In relating research on the effects of negative campaign advertising, we incorporate new research related to independent spending on campaign advertising by Super PACs. Our coverage of research on judicial politics includes research on the increasingly partisan perceptions of Supreme Court decisions. Finally, we changed the example concerning public opinion about US military involvement in foreign affairs; we now refer to recent research on the relationship between public support for war and awareness of disparities in the distribution of the burden of casualties across communities and groups in the United States.

Chapter 2 examines the definition of scientific research and the development of empirical political science. We discuss the role of theory in the research process and review some of the debates in modern and contemporary political science. In response to adopter input, chapter 3 still focuses on the task of helping students to identify and refine appropriate research topics. For adopters who plan to have their students conduct independent research projects, it makes sense to introduce this topic early in the discussion of the research process. This chapter also contains an extended discussion of how to conduct and write a literature review. Chapters 4 and 5 address the building blocks of social scientific research: hypotheses, core concepts, variables, and measurements. Chapter 6 covers research design with an expanded discussion of small-*N* studies, and includes a handy table that summarizes and compares the features of alternative research designs. Chapter 7, on sampling, precedes the chapters on data collection, based on the reasoning that sampling is not used solely by those conducting survey research but also by those using other data collection methods. It also provides important background information for anyone interested in public policy and current events.

Chapters 8 through 10 discuss data collection, with an emphasis on the research methods that political scientists frequently employ and that students are likely to find useful in conducting or evaluating empirical research. We consider the principles of ethical research and the role of human subject review boards and note the ethical issues related to methods of data collection. We examine first-hand observation in chapter 8 and document analysis and the use of aggregate data in chapter 9; the latter now includes an extended example of research based on content analysis. Chapter 10, on survey research, includes a discussion of questionnaire design and tips for face-to-face interviewing. In addition, chapters 9 and 10 also include updated and Web-based sources of aggregate statistics and survey questions and data.

Chapters 11 through 14 focus on data analysis: How do we interpret data and present them to others? All four chapters contain updated examples and discussions. These are supplemented by a host of new figures and tables designed to illustrate the various techniques in as friendly and intuitive a way as possible. We also strengthened the discussion of tests of statistical significance. Chapter 12, on statistical inference, has been streamlined and shortened to focus on the statement and testing of statistical hypotheses and includes examples of significance tests for means and proportions as well as calculating confidence intervals. Our goals are to make the logic of the tests more comprehensible and to stress the differences among statistical, theoretical, and practical significance. In chapter 13, we investigate relationships between two variables. Chapter 14 includes material on logistic regression, an increasingly important statistical tool in social research. In all of this, we attempted to be as rigorous as possible without overwhelming readers with theoretical fine points or computational details. The content is still accessible to anyone with a basic understanding of high school algebra. Our goal, as always, is to provide an intuitive understanding of these sometimes intimidating topics without distorting the concepts or misleading our readers.

Finally, in chapter 15, we present a new research report, using a published journal article that investigates whether satisfaction with life is greater for citizens of countries with larger public sectors or for those who live in countries where the market plays a larger role. This research example ties in well with the research on income inequality used throughout the book to illustrate the research process. As in the past, this article is annotated, although we have changed the format so that students can see more clearly where in the article the authors address key aspects of the research process. We strongly suggest that instructors who assign a research paper have their students consult the example in this chapter and use it to pattern their own writing.

In addition to the "How It's Done" feature, the "Helpful Hints" boxes continue to give students practical tips. Each chapter contains suggested reading lists and lists of terms introduced. A glossary at the end of the book, with more than 250 definitions, lists important terms and provides a convenient study guide.

Companion Web Site: Student and Instructor Resources

The edge every student needs

SAGE edge™
for CQ Press

http://edge.sagepub.com/johnson8e

SAGE edge offers a robust online environment featuring an impressive array of tools and resources for review, study, and further exploration, keeping both instructors and students on the cutting edge of teaching and learning. SAGE edge content is open access and available on demand. Learning and teaching have never been easier!

SAGE edge for Students provides a personalized approach to help students accomplish their coursework goals in an easy-to-use learning environment.

- An online **action plan** includes tips and feedback on progress through the course and materials, which allows students to individualize their learning experience.
- Mobile-friendly **eFlashcards** strengthen understanding of key concepts.
- Mobile-friendly practice **quizzes** encourage self-guided assessment and practice.
- **Chapter summaries** with **learning objectives** reinforce the most important material.

- Carefully selected **Web resources** enhance exploration of key topics.
- **Datasets** and files are available for *Working with Political Science Research Methods, Fourth Edition.*

SAGE edge select for Instructors supports teaching by making it easy to integrate quality content and create a rich learning environment for students.

- **Test banks built on Bloom's Taxonomy** provide a diverse range of test items with Respondus test generation.
- Editable, chapter-specific **PowerPoint® slides** offer flexibility when creating multimedia lectures.
- **Instructor manual** summarizes key concepts to ease preparation for lectures and class discussions.
- A set of all the **graphics from the text** in PowerPoint, PDF, and JPEG formats can be used for class presentations.

Solutions manual is provided for *Working with Political Science Research Methods, Fourth Edition.*

Accompanying Workbook

In addition to updating all of the Web site materials, Jason Mycoff has substantially revised the accompanying workbook, *Working with Political Science Research Methods, Fourth Edition,* providing many new exercises while retaining the ones we feel worked well in the previous edition. Based on user feedback, he looked for opportunities to add more problems for practicing statistical calculations, more variation in subfield coverage, and new datasets. The new edition also includes the student version of SPSS so students can work with their own copy in courses that use SPSS. Each workbook chapter briefly reviews key concepts covered by the corresponding chapter in the text. Students and instructors will find datasets and other documents and materials used in the workbook exercises at http://edge.sagepub.com/johnson8e. The datasets, available on a variety of platforms, may also be used for additional exercises and test items developed by instructors. Instructors may want to add on to the datasets or have their students do so as part of a research project. A solutions manual for adopters of the workbook is also available online at http://edge.sagepub.com/johnson8e.

In closing, we would like to make a comment on statistical software. Instructors remain divided over the extent to which computers should be part of an introductory research course and what particular programs to require. While the student version of SPSS is included with the workbook, neither the workbook exercises nor the textbook problems are written specifically for SPSS. We encourage instructors and students alike to explore the many online statistical resources such as SDA, ICPSR, American Factfinder, Rice Virtual Statistics Lab, and Vassarstats in addition to software like SPSS, STATA, and SAS for their analytical needs.[1]

Acknowledgments

We would like to thank our careful reviewers who helped us shape this new edition: Susan Allen, University of Mississippi; Elisabeth Carter, Keele University; Matthew Childers, University of Georgia; Govinda Clayton, University of Kent; Ashlie Delshad, West Chester University; and Christopher Raymond, Queens University Belfast. Each of these reviewers has helped make the eighth edition stronger than ever, and we are grateful for their assistance.

We would like to thank several people who have contributed to this edition: Talia Greenberg, our copy editor; Olivia Weber-Stenis, our production editor; and Eric Garner, assistant managing editor of production. We are especially thankful for the continued support and patience of acquisitions editor Sarah Calabi and senior development editor Nancy Matuszak at CQ Press. A book with as many bells and whistles as this one needs many sets of eyes to watch over it. We are glad to have had so many good ones.

Janet Buttolph Johnson
H. T. Reynolds
Jason D. Mycoff

[1] Virtually all of the computations in the textbook were performed with R, a statistical programming environment. Although perhaps too difficult for many undergraduates, R is assuming a greater prominence in political science, and we encourage instructors and students to explore this resource, which is totally in the public domain. The program and libraries, documentation, tutorials, books, and other instructional materials can be found on the R Web site: R Development Core Team (2011). R: *A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/

About the Authors

Janet Buttolph Johnson is associate professor of political science and international relations at the University of Delaware, where she specializes in public policy, state and local politics, and environmental policy and politics.

H. T. Reynolds is professor emeritus of political science at the University of Delaware. He is author of *Governing America,* with David Volger; *The Analysis of Nominal Data*, *Second Edition;* and several articles on methodology.

Jason Mycoff is associate professor of political science and international relations at the University of Delaware. His research is on American political institutions, in particular the US Congress, congressional committees, and parties.

CHAPTER 1

Introduction

CHAPTER OBJECTIVES

1.1 Illustrate that political scientists use empirical research methods to investigate important questions about politics and government.

1.2 Discuss the ways that research on inequality examines the importance of group power in determining winners and losers.

1.3 Relate how research attempts to explain why some people participate in politics more than others.

1.4 Summarize findings into what accounts for the gender gap in elected officeholders.

1.5 Explain the approach of studies into the repression of human rights.

1.6 Identify the obstacles of researching judicial decision making.

1.7 Describe the technical issues considered in studying lobbying and oversight of federal agencies.

1.8 Relate the results of studies on the effect of campaign advertising on voters.

1.9 Discuss the findings of research on the factors influencing public support for US military involvement.

POLITICAL SCIENTISTS ARE INTERESTED in learning about and understanding a variety of important political phenomena.

> Some of us are interested in the political differences among countries and wonder why women make up a larger percentage of legislators in some countries than in others, or we may wonder what conditions lead to stable and secure political regimes without civil unrest, rebellion, or government repression.

> Another area of interest is the relationships and interactions between nations and how some nations exercise power over others.
>
> Other political scientists are more interested in the relationship between the populace and public officials in democratic countries and, in particular, whether or not public opinion influences the policy decisions of public officials.
>
> Still others are concerned with how particular political institutions function. Does Congress serve the interests of well-financed groups rather than of the general populace? Do judicial decisions depend upon the personal values of individual judges, the group dynamics of judicial groups, or the relative power of the litigants? To what extent can American presidents influence the actions of federal agencies? Does the use of nonprofit service organizations to deliver public services change government control of and accountability for those services?

These are just a very few examples of the types of questions political scientists investigate through their research.

This book is an introduction to **empirical research**—a methodology that requires scholars to clearly state hypotheses or propositions than can be evaluated with actual, "objective" observation of political phenomena. Students should learn about how political scientists conduct empirical research for three major reasons. First, citizens in contemporary American society are often called upon to evaluate arguments and research about political phenomena. Debates about the wisdom of the death penalty, for example, frequently hinge on whether or not it is an effective deterrent to crime, and debates about term limits for elected officials involve whether or not such limits increase the competitiveness of elections and the responsiveness of elected officials to the electorate. Similarly, evaluating current developments in the regulation of financial markets can be informed by research on what influences the behavior of regulatory agencies and their staff. In these and many other cases, thoughtful and concerned citizens find that they must evaluate the accuracy and adequacy of the theories and research of political (and other social) scientists.

A second reason is that an understanding of empirical research concepts is integrally related to students' assimilation and evaluation of knowledge in their coursework. An important result of understanding the scientific research process is that a student may begin to think more independently about concepts and theories presented in courses and readings. For example, a student might say, "That may be true under the given conditions, but I believe it won't remain true under the following conditions." Or, "If this theory is correct, I would expect to observe the following." Or, "Before I will accept that interpretation, I'd like to have this additional information." Students who can specify what information is needed and what relationships among phenomena must be observed in support of an idea are more likely to develop an understanding of the subjects they study.

A third, and related, reason for learning about political science research methods is that students often need to conduct research of their own, whether for a term paper in an introductory course on American government, a research project in an upper-level seminar, a senior thesis, or a series of assignments in a course devoted to learning empirical research methods. Familiarity with empirical research methods is generally a prerequisite to making this a profitable endeavor.

The prospect of learning empirical research methods is often intimidating to students. Sometimes, students dislike this type of inquiry because it involves numbers and statistics. To understand research well, one must have a basic knowledge of statistics and how to use statistics in analyzing data and reporting research findings. However, the empirical research process that we describe here is first and foremost a way of thinking and a prescription for disciplined reasoning. Statistics will be introduced only after an understanding of the thought process involved in empirical research is established, and then in a way that should be understandable to any student familiar with basic algebra.

Thus, the plan for this book is as follows:

> Chapter 2 discusses what we mean by the scientific study of political phenomena. We also review the historical development of political science as a discipline and introduce alternative perspectives on what is the most appropriate approach to the study of political phenomena; not all political scientists agree that politics can be studied scientifically or that the results of such efforts have been as useful or inclusive of important political phenomena as critics wish.
>
> In chapter 3, we address an aspect of the research process that often poses a significant challenge to students: finding an interesting and appropriate research topic and developing a clearly stated research question. Therefore, in this edition we show how to explore "the literature" and find out what political scientists and others have written about political phenomena in order to sharpen the focus of a research topic, a discussion that came later in previous editions. Chapter 3 focuses on investigating relationships among concepts and developing explanations for political

phenomena. It also includes an example and discussion of how to write the literature review section of a research paper.

Chapter 4 builds on the discussion in chapter 3 by adding the "building blocks" of scientific research: defining complex concepts, hypotheses, variables, and units of analysis.

Chapter 5 addresses the challenge of developing valid and reliable measures of political phenomena. It also discusses how our choices about how we measure variables affect the statistics we may use later to analyze the data we collect.

Chapter 6 presents research designs, both experimental and nonexperimental. The strengths and weaknesses of research designs are discussed, particularly as they relate to causality. The concepts of internal validity and external validity are reviewed here as well.

Chapter 7 covers the logic and basic statistical features of sampling. Various types of samples, including probability and nonprobability samples, are described. Much of our information about political phenomena is based on samples, so an understanding of the strengths and limitations of sampling is important.

Chapter 8 is the first of three chapters discussing the major methods used by political scientists and other social scientists to collect data. This chapter reviews the main methods and the reasons for choosing one method over another. It focuses on observation as a data collection method.

Chapter 9 focuses on the multitude of documents available for use by political scientists, ranging from media clips to diaries to written speeches to the vast body of data collected by government as well as private organizations. It includes a discussion of content analysis, a quantitative approach to analyzing documents.

Chapter 10 discusses interviewing and survey research or polling. It reviews various types of polls and their strengths and weaknesses, as well as the design of survey instruments.

Chapter 11 offers an extensive discussion of descriptive statistics and the analysis of single variables. We present a variety of graphical options useful in displaying data, as visual representations of data are often an extremely effective way to present information. Tips on recognizing and avoiding misleading uses of graphical displays are an essential part of this chapter.

> Chapter 12 is devoted to the concepts of statistical inference, hypothesis testing, and calculating estimates of population parameters. This chapter builds on the foundation established in the earlier chapter on sampling.
>
> Chapter 13 then moves on to the analysis of bivariate data analysis—the investigation of the relationship between two variables.
>
> Chapter 14 is the final statistics chapter. Here we explore statistical techniques used in the quest for explanation and demonstrating causality. These involve multivariate analysis, as the explanation of a political phenomenon rarely is based on simply one other factor or variable.
>
> As in previous editions, we conclude with an annotated example of an actual, peer-reviewed research article. Chapter 15 contains a new example that allows students to see the discussion and application of many of the concepts and statistical procedures covered in earlier chapters.

Researchers conduct empirical studies for two primary reasons. One reason is to accumulate knowledge that will apply to a particular problem in need of a solution or to a condition in need of improvement. Research on the causes of crime, for example, may be useful in reducing crime rates, and research on the reasons for poverty may aid governments in devising successful income maintenance and social welfare policies. Such research is often referred to as **applied research** because it has a fairly direct, immediate application to a real-world situation.

Researchers also conduct empirical research to satisfy their intellectual curiosity about a subject, regardless of whether the research will lead to changes in government policy or private behavior. Many political scientists, for example, study the decision-making processes of voters, not because they are interested in giving practical advice to political candidates but because they want to know if elections give the populace influence over the behavior of elected public officials. Such research is sometimes referred to as **pure, theoretical, or recreational research** to indicate that it is not concerned primarily with practical applications.[1]

Political scientists ordinarily report the results of their research in books or articles published in political science research journals (see chapter 3 for a discussion of how to find articles in these journals). Research reported in academic journals typically contains data and information from which to draw conclusions. It also undergoes peer review, a process by which other scholars evaluate the soundness of the research before it is published. Occasionally, however, political science research

[1] *Recreational research* is a term used by W. Phillips Shively in *The Craft of Political Research,* 2nd ed. (Englewood Cliffs, N.J.: Prentice-Hall, 1980), chap. 1.

questions and analyses appear in newspapers and magazines, which have a wider audience. Such popularly presented investigations may use empirical political science methods and techniques as well.

In the remainder of this chapter, we describe several political science research projects that were designed to produce scientific knowledge about significant political phenomena. We refer to these examples throughout this book to illustrate many aspects of the research process. We present them in some detail now so that you will find the later discussions easier to understand. We do not expect you to master all the details at this time; rather, you should read these examples while keeping in mind that their purpose is to illustrate a variety of research topics and methods of investigation. They also show how decisions about aspects of the research process affect the conclusions that may be drawn about the phenomena under study. And they represent attempts by political scientists to acquire knowledge by building on the research of others to arrive at increasingly complete explanations of political behavior and processes.

Research on Inequality

In 1936 Harold Lasswell published *Politics: Who Gets What, When, How.*[2] Ever since, political scientists have liked this title because it succinctly states an important truth: politics is about winning and losing. No political system, not even a perfectly democratic one, can always be all things to all people. Inevitably, policies favor some and disadvantage others. So important is this observation that one of political science's main tasks is to discover precisely which individuals and groups benefit the most from political struggle and why.

A major controversy in the early years of the twenty-first century has been the apparent growth of economic inequality in the United States. Although there is disagreement among social scientists about the extent of the problem, many now believe that large disparities in income and well-being threaten not just the economy but democracy as well. At times the rhetoric can become feverish:

> The 99.99 percent is lagging far behind. The divide between the haves and have-nots is getting worse really, really fast. . . . If we don't do something to fix the glaring inequities in this economy, the pitchforks are going to come for us. No society can sustain this kind of rising inequality. In fact, there is no example in human history where wealth accumulated like this and the pitchforks didn't eventually come out. You show me a

[2] Harold Lasswell, *Politics: Who Gets What, When, How* (New York: Hittlesey House, 1936). A more recent statement of the idea is found in Benjamin I. Page, *Who Gets What from Government* (Berkeley: University of California Press, 1983).

> highly unequal society, and I will show you a police state. Or an uprising. There are no counterexamples. None. It's not if, it's when.[3]

Other commentators, however, are not as concerned:

> If one looks at after-tax income, the increase in income inequality over time is greatly reduced. If one goes further and factors in the government's attempts to redistribute income, income inequality is not increasing in the U.S. at all. This after-tax, after-transfer income essentially is a measure of how much stuff you can consume (either by buying it or because somebody gave you free stuff). And, as demonstrated by Gary Burtless of The Brookings Institution (a center-left think tank), income inequality measured this way has actually decreased in the U.S. over the decade from 2000–2010.[4]

Inequality has concerned political scientists for decades. Democracy, after all, assumes political equality, and if people have widely varying levels of income, are they (can they be) politically equal? Before reaching definitive conclusions, however, one needs to study systematically and objectively the level, the causes, and the effects of disparities in income and wealth.

In a 2005 study, Lane Kenworthy and Jonas Pontusson analyzed trends in the distribution of gross market income—the distribution of income before taxes and government transfers—for affluent Organisation for Economic Co-operation and Development (OECD) countries using data from the Luxembourg Income Study.[5] Kenworthy and Pontusson were interested in whether inequality in market income had increased and to what extent government policies had responded to changes in market income inequality. In particular, they were interested in testing the median-voter model developed by Allan H. Meltzer and Scott F. Richard.[6]

According to the median-voter model, support for government redistributive spending depends on the distance between the income of the median voter and the average market income of all voters. The greater the average market income is in comparison to the median income, the greater the income inequality and, thus, the greater the

[3] Nick Hanauer, "The Pitchforks Are Coming . . . for Us Plutocrats," *Politico*, June 2014. Accessed December 28, 2014. Available at http://www.politico.com/magazine/story/2014/06/the-pitchforks-are-coming-for-us-plutocrats-108014.html#.VKMNwskO8uc

[4] Jeffrey Dorfman, "Dispelling Myths about Income Inequality," *Forbes*, May 8, 2014. Accessed December 27, 2014. Available at http://www.forbes.com/sites/jeffreydorfman/2014/05/08/dispelling-myths-about-income-inequality/

[5] Lane Kenworthy and Jonas Pontusson, "Rising Inequality and the Politics of Redistribution in Affluent Countries," *Perspectives on Politics* 3, no. 3 (2005): 449–71. Available at http://www.u.arizona.edu/~lkenwor/pop2005.pdf

[6] Allan H. Meltzer and Scott F. Richard, "A Rational Theory of the Size of Government," *Journal of Political Economy* 89, no. 5 (1981): 914–27.

demand from voters for government spending to reduce this gap. Countries with the greatest market inequalities should have more such government spending.

One way to test the median-voter model is to see whether changes in redistribution are related to changes in market inequality. One would expect that larger changes in market inequality would cause larger changes in redistribution if governments are responsive to the median voter. Kenworthy and Pontusson found this to be the case, although the United States, Germany, and the United Kingdom did not fit the pattern very well. In further analyses in which they looked at country-by-country responsiveness to market inequality over several decades, they found that most OECD countries are responsive to market income inequalities, although to varying degrees, and that the United States is the least responsive.

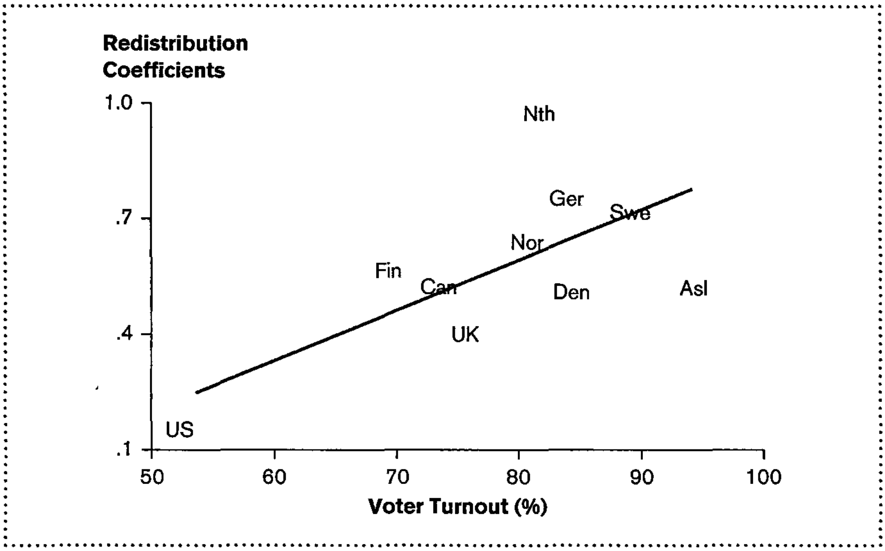
Perhaps, Kenworthy and Pontusson suggested, government responsiveness to market inequality is related to voter turnout. If one assumes that lower-income voters are less likely to turn out to vote than are higher-income voters, then one would expect that the lower the turnout, the less likely governments would be pressured to respond to income inequality. The median-voter model still would apply, but in countries with low voter turnout, the median voter would be less likely to represent lower-income households. Kenworthy and Pontusson used regression analysis and a scatterplot (you will learn about these in chapter 12), shown in figure 1-1, to show that the higher the voter turnout, the more responsive a country is to market income inequality. The results provide a possible explanation for why the United States is less responsive to changes in market inequality than are other nations: the United States has the lowest turnout rate among the nations included in the analysis.

These research findings are interesting in view of another body of research that we will consider shortly: the possibility that since the mid-1950s, the bottom classes in America have been increasingly dropping out of electoral politics.

In 2010 an entire issue of the journal *Politics & Society* was devoted to the topic of income inequality. In the lead article, "Winner-Take-All Politics: Public Policy, Political Organization, and the Precipitous Rise of Top Incomes in the United States," Jacob S. Hacker and Paul Pierson took issue with much of the previous research on the causes of income inequality in the United States.[7] First, they dismissed economic accounts that attribute growth in inequality to "apolitical processes of economic change" for failing to explain differences among nations, as illustrated in figure 1-2. This figure shows that the top 1 percent's share of national income is the highest in the United States (16%) and that it increased the most,

[7] Jacob S. Hacker and Paul Pierson, "Winner-Take-All Politics: Public Policy, Political Organization, and the Precipitous Rise of Top Incomes in the United States," *Politics & Society* 38, no. 2 (2010): 152–204.

FIGURE 1-1 **Redistribution Coefficients by Average Voter Turnout**



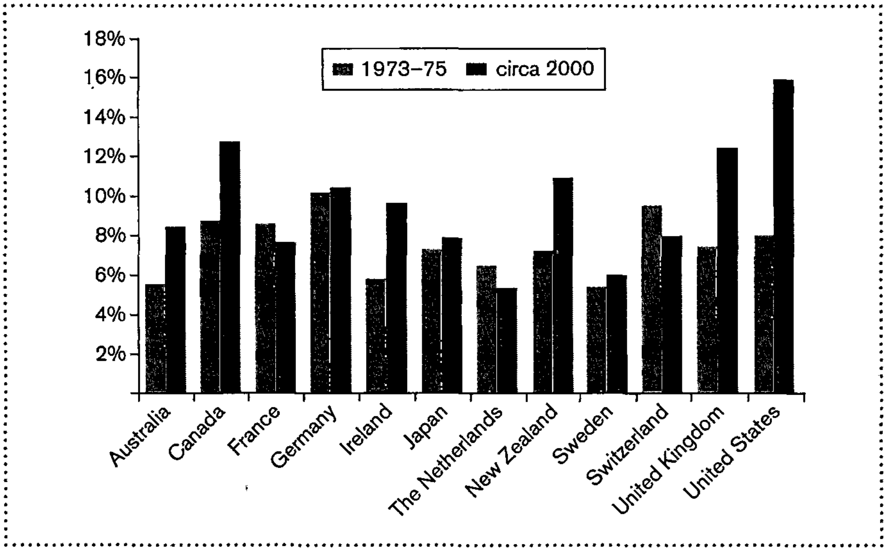
Source: Reprinted from Lane Kenworthy and Jonas Pontusson, "Rising Inequality and the Politics of Redistribution in Affluent Countries," *Perspectives on Politics* 3, no. 3 (2005): 462.

Note: Asl = Australia; Can = Canada; Den = Denmark; Fin = Finland; Ger = Germany; Nth = The Netherlands; Swe = Sweden; UK = United Kingdom; US = United States. Presidential elections for the United States; general parliamentary elections for the other countries. Redistribution data are for working-age households only.

almost doubling, between the 1970s and 2000. Second, they attacked previous political analyses on three counts: for downplaying "the extreme concentration of income gains at the top of the income ladder" (figure 1-3 shows the gain in the top 1 percent's share of national pretax income from 1960 to 2007), for missing the important role of government policy in creating what they called a "winner-take-all" pattern, and for focusing on the median-voter model and electoral politics instead of important changes in the political organization of economic interests. They argued that the median-voter model and the extreme skew in income don't add up. Even accounting for lower turnout among lower income voters, the difference between the income of the median voter and the incomes at the very top is too big to argue that politicians are responding to the economic interests of the median voter.

Their explanation for the "precipitous rise" in top incomes in the United States rejects the median-voter model. Instead, they argue that policies governing corporate structure and pay, the functioning of financial markets, and the framework of

FIGURE 1-2 **The Top 1 Percent's Share of National Income, Mid-1970s versus Circa 2000**



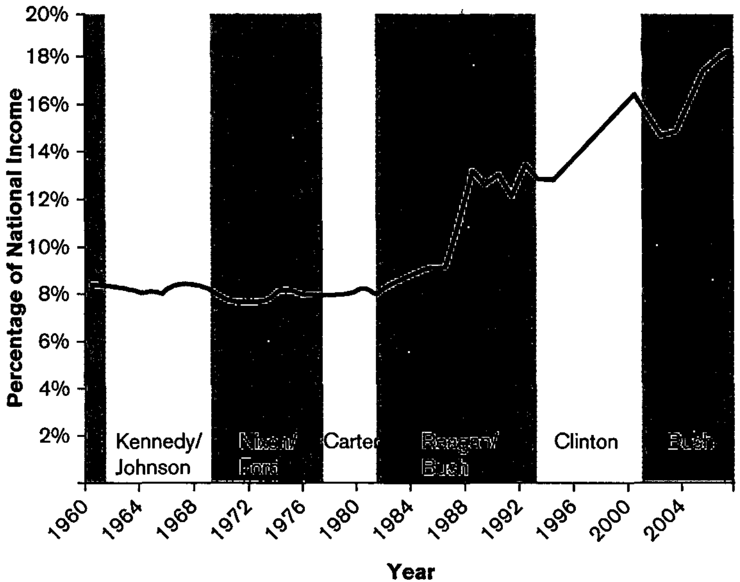
Source: Andrew Leigh, "How Closely Do Top Incomes Track Other Measures of Inequality?" *Economic Journal* 117, no. 524 (2007): 619–33, http://econrsss.anu.edu.au/~aleigh/pdf/ TopIncomesPanel.xls, cited in Jacob S. Hacker and Paul Pierson, "Winner-Take-All Politics: Public Policy, Political Organization, and the Precipitous Rise of Top Incomes in the United States," *Politics & Society* 38, no. 2 (2010): fig. 2, p. 160. Copyright © 2010 SAGE Publications. Reprinted by Permission of SAGE Publications.

industrial relations have had much to do with changes in pretax income (so-called market income).

More recent research on winners and losers in politics has focused on the impact of economic inequality on political representation in government, whether the Republican and Democratic Parties differ in their response to wealthy constituents, and whether minorities fare better under Democratic rather than Republican administrations. In his research, Thomas J. Hayes examines Senate responsiveness to economic constituencies for the 107th through the 111th Congresses.[8] Hayes had to overcome a common problem encountered by political scientists who want to study how closely the voting records of members of Congress match public opinion in their districts—quite often national opinion polls do not contain enough respondents from every district to gauge district public opinion accurately.

[8] Thomas J. Hayes, "Responsiveness in an Era of Inequality: The Case of the U.S. Senate," *Political Research Quarterly* 63, no. 3 (2012): 585–99.

FIGURE 1-3 **The Richest 1 Percent's Share of National Pretax Income, 1960–2007**



Sources: Thomas Piketty and Emmanual Saez, "Income Inequality in the United States, 1913–1998," *Quarterly Journal of Economics* 118, no. 1 (2003): 1–39; updated tables and figures for Piketty and Saez available at http://elsa.berkeley.edu/~saez/TabFig2007.xls, as cited in Jacob S. Hacker and Paul Pierson, "Winner-Take-All Politics: Public Policy, Political Organization, and the Precipitous Rise of Top Incomes in the United States," *Politics & Society* 38, no. 2 (2010): fig. 1, p. 156. Copyright © 2010 SAGE Publications. Reprinted by Permission of SAGE Publications.

Note: Excluding capital gains.

(You will learn about sampling accuracy in chapter 7.) Luckily for Hayes, the 2004 National Annenberg Election Survey (NAES) included enough respondents from every state. Also fortunately for Hayes, Republicans had unified control of the national government for most of the 107th through 109th Congresses, while Democrats controlled the Senate in the 110th and 111th Congresses with unified control for the 111th, allowing him to compare the responsiveness of senators under periods of different party control of the Senate and its agenda. Hayes divided NAES respondents into terciles based on income. He measured public opinion using respondents' self-placement on an ideology scale ranging from -2 to 2 with lower values coded as liberal and calculated the average ideology score for each of the three income groups. For senators, he used a measure designed to summarize legislators' ideological positions based on all the votes they cast in each Congress. Hayes then used regression analysis to see how senators' voting records matched up with the positions of each of the three income groups. Hayes concludes that "Senators do not respond to the views of all their constituents in an equal manner."[9] Instead, he found significant responsiveness toward upper-income constituents in each Congress. Regardless of which party controlled the Senate, he was unable to detect responsiveness toward low-income constituents. He also found that Republicans were more responsive than Democrats to middle-income constituents in the 109th Congress, and in the 107th Senate, responsiveness toward the upper-income constituents increased once Democrats took control of the chamber, contrary to expectations that Democrats are less responsive to upper-income constituents than are Republicans. As interesting as these findings may be, they are limited in the extent to which they can be generalized to the US Senate in other time periods or to the US House of Representatives. Our last example of research on "who gets what" in politics uses data that cover a considerably longer period of time.

Zoltan L. Hajinal and Jeremy D. Horowitz's research on racial winners and losers addresses the question of whether there is a difference between the Republican and Democratic Parties in the responsiveness to different constituencies, in this case to minorities.[10] They note that Democratic Party leadership has argued that liberal policies in race, welfare, education, crime, and other social issues lead to better outcomes for minorities, while the Republican Party argues that policies that help the economy in general and reduce the size of government lead to higher growth and higher incomes for all, including minorities, and provide a greater benefit than that supplied by specific federal programs. Hajinal and Horowitz set out to examine the evidence to see if it fits one claim over the other by looking at the correlation between party control and minority well-being.[11] They decided to measure party control by which party controls the presidency (although they also checked to see if party control of Congress, the median ideology of the Supreme Court, and percentage of US Court of Appeals judges nominated by a Democratic president are important). They measured minority well-being in terms of black median family income, unemployment, and poverty in some tests, while in others they used criminal justice, educational attainment, and health indicators. They also took into account differences in the condition of the national economy and the likelihood that effects lag behind periods of party control. Their analysis concluded that the black population fares better under Democratic presidents, with black family income growing over $1,000 faster annually, poverty rates declining

[9] Ibid., 595.

[10] Zoltan L. Hajinal and Jeremy D. Horowitz, "Racial Winners and Losers in American Party Politics," *Perspectives on Politics* 12, no. 1 (2014): 100–118.

[11] Zoltan L. Hajinal and Jeremy D. Horowitz, "Racial Winners and Losers in American Party Politics," *Perspectives on Politics* 12, no. 1 (2014): 100–18.

2.6 points faster, and unemployment rates falling almost one point faster than under Republican administrations.[12] Hajinal and Horowitz also checked to see if black gains came at the expense of whites. They found that whites also made gains under Democratic administrations, but that the difference in gains for whites under Democratic administrations as compared to Republican administrations was much smaller than for blacks and were not statistically significant. But Hajinal and Horowitz noted that more research needs to be done to determine which policies account for minority gains.

The point of this example is not to make a statement about the value of particular ideologies or parties. Instead, we want to stress that important questions—what could be more crucial than knowing who gets what from a political system?—can be answered systematically and objectively, even if tentatively, through careful thought and analysis. Among other things, this research demonstrates how political scientists must take advantage of naturally occurring changes in our political

[12] Their results are shown in the following table:

Party of the President and Annual Change in Black Economic Well-Being: Regression

	Average Annual Change for Blacks			Average Gain of Blacks Relative to Whites		
	Income	Poverty	Unemployed	Income	Poverty	Unemployed
Democratic president	1031 (276)***	–2.61 (.85)**	–.87 (.38)*	403 (193)*	–2.41 (.82)**	–.57 (.24)*
Median income	.071 (.054)	.000 (.000)	.000 (.000)	.007 (.038)	.000 (.000)	.000 (.000)
Inflation	–213 (63)***	.48 (.18)*	.24 (.09)*	–54.4 (44.1)	.22 (.18)	.14 (.06)*
Change in labor force	753 (518)	–1.59 (1.75)	–2.10 (.69)**	–31.4 (363)	–.60 (1.69)	–1.02 (.43)*
Change in oil prices	.24 (6.07)	.002 (.016)	.011 (.008)	5.80 (4.26)	–.005 (.015)	.001 (.005)
Time trend	–2110 (2054)	–.281 (6.32)	–1.20 (2.90)	–865 (1442)	–2.89 (6.12)	–1.15 (1.82)
Democratic house	974 (500)	1.80 (1.90)	–.44 (.70)	921 (351)*	.63 (1.84)	–.31 (,44)
Democratic senate	–240 (388)	–.26 (1.06)	.10 (.50)	–489 (272)	.67 (1.03)	–.13 (.31)
Constant	–2273 (1523)	–22.5 (11.6)	.34 (2.61)	–729 (1070)	–22.7 (11.3)	.13 (1.64)
Adj R. squared N	.29 57	.39 40	.35 51	.19 57	.29 40	.19 51

***p<.001**p<.01*p<.05 standard error in parentheses

Ibid., 106

system when they happen, and how their findings contribute to our ability to evaluate the performance of elected officials. Moreover, we hope to show that the techniques used in these debates are not beyond the understanding of students of the social sciences.

Who Votes? Who Doesn't?

The previous example of research showed the importance of group power in determining political winners and losers. Political participation is a major factor: those individuals who make themselves heard in politics "do better" than do those who are apathetic. So a natural question is, Why do some people participate more than others?

A good place to start looking for the answer is with the decision to vote. Except for new research, which we review briefly later in this chapter, most political scientists accept two generalizations about voting in the United States. First, voting varies by socioeconomic class. Members of the lower classes participate less frequently than do more affluent and better-educated citizens. There is, in short, a "class gap" in turnout rates.[13] The second finding is that since the 1950s, a smaller and smaller proportion of the population has been going to the polls. Voting rates in federal elections have dropped more or less steadily, rebounding somewhat to 55 percent in 1992 but falling to a new low of less than 50 percent in 1996. Since then, turnout has increased in presidential elections to 56.8 percent in 2008. The voting rate has been even lower in recent congressional or "off-year" elections and in the South.

The political scientist Walter Dean Burnham combined these findings into an argument that has come to be known as "selective class demobilization."[14] In a nutshell, Burnham's thesis is that the decline in turnout is especially pronounced among those in the lower and working classes; those with relatively little education and income; and those who work in manual, routine service, and unskilled occupations. Those higher up the ladder, so to speak, have voted at more or less the same rates since the 1950s. In Burnham's words, "The attrition rate among various working-class categories is more than three times as high as in the professional and technical category and well over twice as high as for the middle class as a whole."[15] In other words, for every upper-class nonvoter, there are now two or three lower-class nonvoters. It appeared to Burnham and others that the lower classes are effectively abandoning electoral politics. As a consequence, and in keeping with the research on winners and losers, it appears that political rewards may increasingly favor the middle and upper classes.

[13] Thomas E. Patterson, *The Vanishing Voter* (New York: Vintage Books, 2003), 44–46.

[14] Walter Dean Burnham, "The Turnout Problem," in A. James Richley, ed., *Elections American Style* (Washington, D.C.: Brookings Institution, 1987).

[15] Ibid., 125.

In the tradition of modern political science, Burnham supported his case by using hard empirical data—measured turnout rates for various social strata. (He relied heavily on census data and graphical and tabular displays to make his points.) But even more important, he supplied a theory that explains this apparent selective demobilization. He contended that political parties in America, never very strong to begin with, have become even weaker in the post–World War II years as a result of many factors, including the rise of candidate-centered campaigns and the increased use of primary elections in party nominations. The weakening of party organization has been especially pronounced in the Democratic Party.[16] The decline of parties places an especially onerous burden on the working and lower classes. Why? Because these groups, having less education and information about government, rely more heavily on cues and motivation supplied by political parties; without this guidance, these citizens lose their way in politics and frequently drop out.[17]

So selective class demobilization has a cause (the decline of parties) and a consequence (the loss of political influence). If true, Burnham's analysis would have enormous implications for the understanding of American politics. Stated bluntly, public policy will have an upper-class bias. Being so provocative, Burnham's thesis naturally sparked considerable comment and controversy, a fact that illustrates an important aspect of scientific research.

As discussed in chapter 2, science demands independent verification of findings. Conclusions such as Burnham's are not accepted at face value but must be verified by others working separately. In this case, additional research has produced mixed results. Some investigators agree with Burnham that the decline in turnout has been concentrated disproportionately among lower socioeconomic classes.[18] Some have investigated alternative explanations for a decline in turnout among lower socioeconomic classes. In research that has great relevance in light of the research mentioned earlier on income inequality, Frederick Solt investigated the "Schattschneider hypothesis," named after the political scientist E. E. Schattschneider.[19] Schattschneider suggested in 1960 that low participation and high-income voter bias are the result of economic inequality because as the rich grow richer relative to other citizens, they also grow better able to define the alternatives that are considered within the political system and exclude matters of importance to

[16] Burnham wrote, "While no one doubts that the Republican party suffers from some internal divisions and even occasional bouts of selective abstention among its supporters . . . the GOP remains much closer to being a true party in the comparative sense than do today's Democrats" (ibid., 124). This remark is as true in the early twenty-first century as it was in the mid-1980s, when Burnham wrote it.

[17] Ibid., 123–24.

[18] Stephen E. Bennett, "Left Behind: Exploring Declining Turnout among Noncollege Young Whites, 1964–1988," *Social Science Quarterly* 72, no. 2 (1991): 314–33; and Patterson, *The Vanishing Voter*, chap. 2.

[19] Frederick Solt, "Does Economic Inequality Depress Electoral Participation? Testing the Schattschneider Hypothesis," *Political Behavior* 32, no. 1 (2010): 285–301.

poor citizens.[20] Solt found that citizens of states with greater income inequality are less likely to vote in gubernatorial elections and that income inequality increases income bias in the electorate, thus providing empirical support for Schattschneider's hypothesis. But others, using alternative measures of class and other data sets, have come to a conclusion different from Burnham. Jan E. Leighley and Jonathan Nagler, for instance, found "that the class bias [in nonvoting] has not increased since 1964."[21]

Complicating matters further, recent research calls into question even the basic belief that voter turnout in general has been declining. These newer investigations say the apparent decrease in the rate of electoral participation stems from an artifact in how turnout is measured. The voting rate has typically been measured as the number of votes cast divided by the number of eligible voters. This procedure may seem straightforward, but a problem arises: How should the eligible voting population be defined? The Census Bureau uses the so-called voting-age population (VAP) as its measure of the eligible electorate. But, as Michael P. McDonald and Samuel L. Popkin maintained, this approach "includes people who are ineligible to vote, such as noncitizens, felons, and the mentally incompetent, and fails to include [Americans] living overseas but otherwise eligible."[22] They developed an alternative measure of the pool of legally eligible voters or voting-eligible population (VEP) and showed that when it is used in the denominator of voting-rate calculations, "nationally and outside the South there are virtually no identifiable turnout trends from 1972 onward, and within the South there is a clear trend of *increasing* turnout rates [emphasis added]."[23]

In a 2010 article, "Does Measurement Matter? The Case of VAP and VEP in Models of Voter Turnout in the United States," Thomas Holbrook and Brianne Heidbreder investigated the impact of using VAP or VEP on our understanding of the causes of variation in turnout among states.[24] Because states control numerous factors affecting the ease of voting (such as early voting, voting by absentee ballot, and

[20] E. E. Schattschneider, *The Semisovereign People: A Realist's View of Democracy in America* (New York: Holt, Reinhart, and Winston, 1960).

[21] Jan E. Leighley and Jonathan Nagler, "Socioeconomic Class Bias in Turnout, 1964–1988: The Voters Remain the Same," *American Political Science Review* 86, no. 3 (1992): 734. Also see Ruy A. Teixeria, *Why Americans Don't Vote: Turnout Decline in the United States, 1960–1984* (Westport, Conn.: Greenwood, 1987).

[22] Michael P. McDonald and Samuel L. Popkin, "The Myth of the Vanishing Voter," *American Political Science Review* 95, no. 4 (2001): 963. Available at http://elections.gmu.edu/APSR McDonald and_ Popkin_2001.pdf

[23] Ibid., 968. Also see Michael P. McDonald, "On the Overreport Bias of the National Election Study Turnout Rate," *Political Analysis* 11, no. 2 (2003): 180–86.

[24] Thomas Holbrook and Brianne Heidbreder, "Does Measurement Matter? The Case of VAP and VEP in Models of Voter Turnout in the United States," *State Politics & Policy Quarterly* 10, no. 2 (2010): 159–81.

variable registration deadlines), whether or not gubernatorial elections are held concurrently with federal elections, and whether ballot initiatives are allowed or not, studying voter turnout at the state level can tell us a lot about the relative importance of these factors and other determinants of turnout. Considerable variation exists among the states regarding voting restrictions placed on felons and the size of the noncitizen population. Therefore, the difference between VAP and VEP for some states could be significant. Using VAP as the measure of turnout could mask the impact of factors on the turnout of those voters actually eligible to vote. Holbrook and Heidbreder's analysis showed a strong correlation exists between VEP and VAP, but for some states the two measures do diverge. Furthermore, using VEP rather than VAP changes the extent to which per capita income, Hispanic population, and number of ballot initiatives are found to affect voter turnout: per capita income and ballot initiatives become more significant and Hispanic population less so when VEP is used.

Adjusting VAP to exclude felons raises another series of questions investigated by political scientists: Why is the United States alone among democratic countries in this regard? To what extent does the disenfranchisement of nonincarcerated offenders (a practice that in the United States results in the disenfranchisement of large numbers of citizens) alter the outcome of elections? And what accounts for differences in restricting access to the ballot among the American states?[25]

Finally, here is another curious twist in research on voter turnout. Some investigators approach the study of political phenomena by building what are known as formal models. Modelers begin with a set of a priori assumptions and propositions and use logic to deduce further statements from them. In the case of voter turnout, the modeling approach begins with the assumption that citizens are rational, in the sense that they try to maximize their utility (the things that they value) at the least cost to themselves. So a potential voter will think about the personal benefits of going to the polls and weigh these against the costs of doing so (for example, taking the time to become informed, registering, and finding and driving to the polling place). Surprisingly, many models lead to the conclusion that a rational person—one who wants to maximize utility at least cost—will decide that voting is not worth the effort and simply abstain.[26] The reason for this conclusion: one single individual's participation has an exceedingly small probability of affecting the outcome of an election. So, according to the deduction, the small chance of bringing benefits by voting is easily outweighed by the costs, however low. Consequently,

[25] Jeff Manza and Christopher Uggen, "Punishment and Democracy: Disenfranchisement of Nonincarcerated Felons in the United States," *Perspectives on Politics* 2, no. 3 (2004): 491–505. Available at http://www.soc.umn.edu/~uggen/Manza_Uggen_POP_04.pdf

[26] One of the first to arrive at this conclusion was the economist Anthony Downs, whose seminal book *An Economic Theory of Democracy* (New York: Harper and Row, 1957) sparked a generation of research into the seeming irrationality of voting.

the formal model predicts that hardly anyone will vote. But in point of fact millions of Americans do vote, which seems to belie the model's conclusions. This situation, which has been called the "paradox of voting," has sparked an enormous amount of discussion and controversy since the 1950s.[27] One recent attempt to explain why citizens in a democracy vote, therefore, includes psychological or motivational variables in a model of voting.[28] This is not the place to sort out all of the research related to voter turnout. Instead, we have used studies of voter turnout to illustrate some features of research that are described in more detail in the following chapters, including the derivation of hypotheses from existing theory, measurement of concepts, and the use of objective standards to adjudicate among competing ideas. The major point, perhaps, is that if one's procedures are stated clearly, others can pick up the thread of analysis and independently investigate the problem. In this sense empirical political research is, like all science, a cumulative process. Usually no one person or group can discover a definitive answer to a complicated phenomenon like voting or nonvoting. Rather, the answers come—if they do at all—from the gradual accumulation of findings from numerous investigators working independently of one another to validate or invalidate each other's claims. Finally, research on voter turnout shows the connections among empirical research, values, and public policy. People look at research to obtain useful knowledge about voter turnout rates as a whole or differences in voter turnout rates for different groups of potentially eligible voters. Those who believe that turnout rates should be higher may advocate changes in public policy to encourage voting.

Politics and the Gender Gap

Much has been written about underrepresentation of women in public office. Based on data from 187 countries, women made up on average only 16.60 percent of the legislators in the lower house of parliament in 2008.[29] Rwanda had the highest percentage, with 48.8 percent. The United States was at the average in 2008, at 16.8 percent. After the elections in 2014, numbers in the United States increased only slightly. In the 114th US Congress there are 104 women, only 18 percent of the membership, although this is an increase from 99 women in the previous Congress.

[27] See Donald P. Green and Ian Shapiro, *The Pathologies of Rational Choice Theory: A Critique of Applications in the Social Sciences* (New Haven, Conn.: Yale University Press, 1994); and Jeffrey Friedman, ed., *The Rational Choice Controversy: Economic Models of Politics Reconsidered* (New Haven, Conn.: Yale University Press, 1996).

[28] Joshua Harder and Jon A. Krosnick, "Why Do People Vote? A Psychological Analysis of the Causes of Voter Turnout," *Journal of Social Issues* 64, no. 3 (2008): 525–49.

[29] Calculation by the author based on data obtained from Democracy Crossnational Data, Release 3.0 (Spring 2009). Accessed January 24, 2015. Available at http://www.hks.harvard.edu/fs/pnorris/Data/Data.htm

Women make up 19.3 percent of the House of Representatives.[30] At the state level the average of female state legislators is 24.3 percent, but the picture is quite varied, with Colorado and Vermont at the top with 42 percent and 41 percent, respectively, and Oklahoma and Louisiana at the bottom with 12.8 percent and 12.5 percent, respectively.[31] What accounts for this gender gap? Is it because women make up a small proportion of the professions that are typical recruiting grounds for candidates? Are women less interested in politics and running for office, and if so, why? Do family considerations weigh more heavily on women, making the demands of public office too difficult to contemplate?

Research by Richard L. Fox and Jennifer L. Lawless addresses these questions. In a national random sample of nearly four thousand high school and college students, they found "a dramatic gender gap in political ambition."[32] In looking for explanations for this gender gap, they found that parental encouragement, politicized educational and peer experiences, participation in competitive activities, and a sense of self-confidence are associated with a young person's interest in running for public office, but that young women report less of these factors than young men and that the gap between men and women in college is greater than in high school. In other research, Fox and Lawless study the political ambitions of men and women in professions (lawyers, business leaders, educators, and political activists) typically thought of as recruitment grounds for candidates for public office. Even though they found a "deeply gendered distribution of household labor and child care among potential candidates," they deemed that differences in family roles and responsibilities did not account for lower levels of political ambition reported by women. Even women unencumbered by family responsibilities reported less political ambition than men. They conclude that candidate recruitment and self-perceived qualifications are the best explanations for the gender gap in political ambition. Women are less likely than men to report that they have been recruited to run for public office by a party leader, elected official, or political activist, or to consider themselves qualified to run for public office even after controlling for differences in family structures, roles, and responsibilities.[33]

What happens when women are elected to political office? What is the effect of the presence of women in legislative bodies? Does it result in substantive as well as

[30] "2014: Not a Landmark Year for Women, Despite Some Notable Firsts," Center for American Women and Politics, Rutgers University. Accessed January 24, 2015. Available at http://www.cawp.rutgers.edu/press_room/news/documents/PressRelease_11-05-14-electionresults.pdf

[31] "Women in State Legislatures 2015: Numbers Still Suck," Center for American Women and Politics, Rutgers University. Accessed January 24, 2015. Available at http://www.cawp.rutgers.edu/press_room/news/documents/PressRelease_01-06-15_stleg.pdf

[32] Richard L. Fox and Jennifer L. Lawless, "Uncovering the Origins of the Gender Gap in Political Ambition," *American Political Science Review* 108, no. 3 (2014): 499–519.

[33] Richard L. Fox and Jennifer L. Lawless, "Reconciling Family Roles with Political Ambition: The New Normal for Women in Twenty-First Century U.S. Politics," *The Journal of Politics* 76, no. 2 (2014): 398–414.

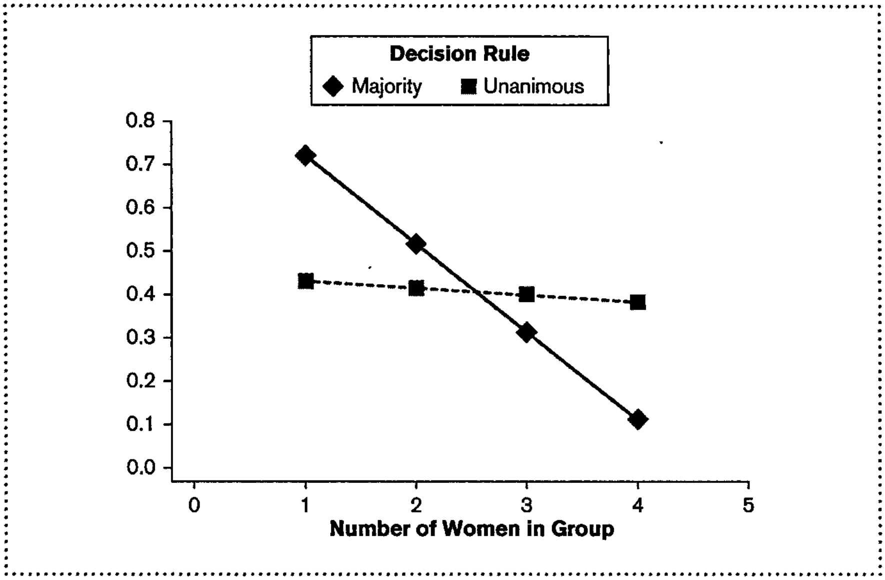
symbolic representation (the perception that women can and should govern)? Is a "critical mass" necessary before such representation effects occur? Is the number of women in a legislative body the critical factor, or might the rules governing deliberation in the legislature also be important? This latter factor is one investigated by Tali Mendelberg, Christopher F. Karpowitz, and J. Baxter Oliphant.[34] They note that research has not shown a clear, positive effect of descriptive representation (number or proportion of women) for women's substantive or symbolic representation. They propose that "the way in which participants interact while speaking may enhance or undermine women's status in deliberation, and that numbers affect this interaction, but in combination with rules." In particular, they note that previous research on the "authoritative use of speech acts" indicates that men are more likely to speak first and talk longer, receive positive feedback on their input, interrupt others in a negative manner, and fail to yield when interrupted. Women tend to speak less and not in the beginning of deliberations, receive little or no positive feedback on their ideas, be interrupted in a negative manner, and yield when interrupted.

Mendelberg, Karpowitz, and Oliphant's research investigates whether these patterns are affected by a group's decision rule: by majority or by consensus or unanimity. They hypothesize that under a unanimous rule, women will receive more respect in deliberations and the expectation of deference by women during discussions will be overridden, *but* only when women are in the minority, not when they predominate (based on previous research). To test their hypothesis, they set up 94 five-member discussion groups composed of between 0 and 5 women, and randomly assigned each group to unanimous or majority rule. Each group was given the identical decision task except for the decision rule. The researchers recorded and transcribed each individual's speech. They counted the number of times each person spoke and coded the number and tone (positive, neutral, or negative) of interruptions, the gender of the speaker, and the gender of the person interrupting.

The following figure shows just some of the results. Graphical representation of data is an efficient and effective way of presenting research findings, and learning how to interpret such graphs is an important, albeit at times challenging, aspect of reading research articles. Figure 1-4 shows the negative proportion of negative and positive interruptions (neutral interruptions are not included in this analysis) received by women from men by group decision rule and number of women in the group. The proportion of negative interruptions is measured on the vertical axis, the number of women in the group is measured along the horizontal axis, and each line represents the type of decision rule. In majority-rule groups, the composition of

[34] Tali Mendelberg, Christopher F. Karpowitz, and J. Baxter Oliphant, "Gender Inequality in Deliberation: Unpacking the Black Box of Interaction," *Perspectives on Politics* 12, no. 1 (2014): 18–44.

FIGURE 1-4 **Negative Proportion of Negativé and Positive Interruptions Received by Women from Men by Group Decision Rule**

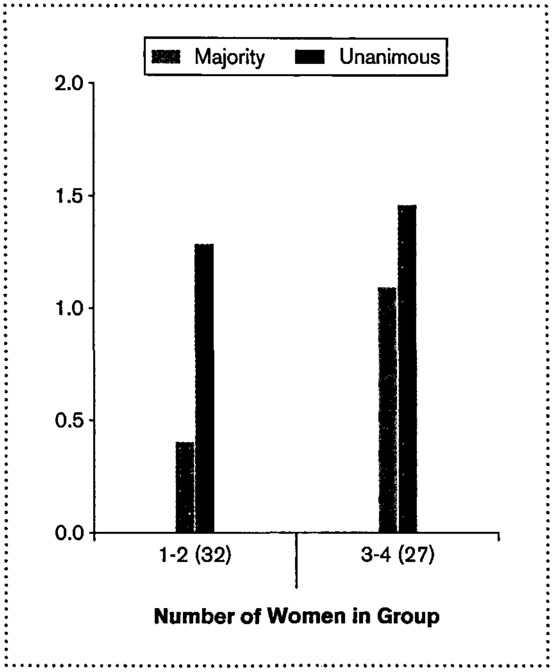


Source: Tali Mendelberg, Christopher F. Karpowitz, and J. Baxter Oliphant, "Gender Inequality in Deliberation: Unpacking the Black Box of Interaction," *Perspectives on Politics* 12, no. 1 (2014): fig. 1, p. 24.

the group has a clear effect on the proportion of negative comments, ranging from over 70 percent when there is only one woman in the group to less than 20 percent when there are four women. Under unanimous rule, the tone of interruptions women receive from men is positive (less than half of the interruptions are negative) and the number of women in a group has no effect on the proportion of negative interruptions. Compared to majority rule, the unanimous rule helps women when they are in the minority; when women are in the minority in majority-rule groups the tone of interruptions they receive from men is negative. But when women are in the majority in majority-rule groups (and their votes are necessary to win), the tone of men's interruptions becomes positive. In this decision-making context, women's status is important, and they are afforded more respect.

Another way of looking at the gender gap in deliberation is to compare men and women with respect to the relative frequency with which they receive positive interruptions. Relative frequencies, a data analysis technique described at length in chapter 11, are a common method of summarizing data. To make the graph in figure 1-5, for every mixed-gender group the researchers take the proportion of a

FIGURE 1-5 **Ratio of Women's to Men's Positively Interrupted Speaking Turns, Mixed Groups (Raw)**



Source: Tali Mendelberg, Christopher F. Karpowitz, and J. Baxter Oliphant, "Gender Inequality in Deliberation: Unpacking the Black Box of Interaction," *Perspectives on Politics* 12, no. 1 (2014): fig. 1, p. 24.

person's speaking turns that received a positive interruption, and calculate the group's average for women divided by its average for men. Next, they separate the groups by decision rule and average the results for groups in which women are in the minority and for groups in which they are in the majority. One can see in figure 1-5 that women receive less than half of the proportion of positive interruptions as men (the horizontal red line represents equal proportions) when they are in the minority in majority-rule groups. In other decision-making contexts women receive about the same or even higher proportions of positive interruptions than men.

This research by Mendelberg, Karpowitz, and Oliphant makes an important contribution to understanding links between demographic representation and substantive and symbolic representation of women, as well as to the broader question of under what circumstances participation in group deliberations by low-status individuals leads to their voices being heard. Clearly, research by political scientists on the gender gap in politics gives us plenty to think about and many questions yet to be answered.

Repression of Human Rights

As a result of improvements in the availability of data, public and scholarly interest focused on the human rights practices of governments has increased substantially over the past two decades. Several organizations (Amnesty International, the US Department of State, and Freedom House, for example) publish annual reports on the human rights performance of nations worldwide. More recently, information and news about human rights has become available on the Internet.

Much of the research on human rights has tried to explain cross-national variation in protection and enjoyment of three legally recognized human entitlements: security rights or personal integrity rights, which include the rights to be free from arbitrary or politically motivated torture, execution, and imprisonment; subsistence rights or basic human needs; and civil and political liberties, which include political and economic rights. They also noted there has been considerable discussion about the relationships among these rights, particularly whether all human rights are indivisible and interdependent (I/I) or whether the protection of basic rights—security and subsistence rights—is distinct and necessary for the enjoyment of all other rights (social, cultural, and economic), or whether there are trade-offs between types of rights—that governments can provide more of one if they restrict another. Research into human rights illustrates the importance of accurate measurement of concepts, in this case the measurement of the various types of rights. We will have more to say about this in chapter 5, but, as an example, researchers Wesley T. Milner, Stephen C. Poe, and David Leblang discussed two approaches to the measurement of personal integrity abuses in their study on trends in human rights and linkages between types of rights.[35] One is events-based and relies on newspaper accounts. This poses a problem in that research typically uses Western newspapers, which may not report abuses systematically and without bias. Furthermore, especially closed regimes may prevent abuses from appearing in news reports. An alternative approach, the standards-based approach, involves coders reading various reports on governments' human rights practices and classifying countries according to a set of predetermined criteria. This approach, too, has its problems, but it is more likely to result in relatively accurate measures for comparison across nations. Researchers Laura Minkler and Shawna Sweeney faced many choices of human rights indicators and some challenges in creating a composite indicator that measured the extent to which countries protected security and subsistence rights simultaneously.[36] Measurement matters, and they spent a considerable amount of time explaining the measures and justifying their choices.

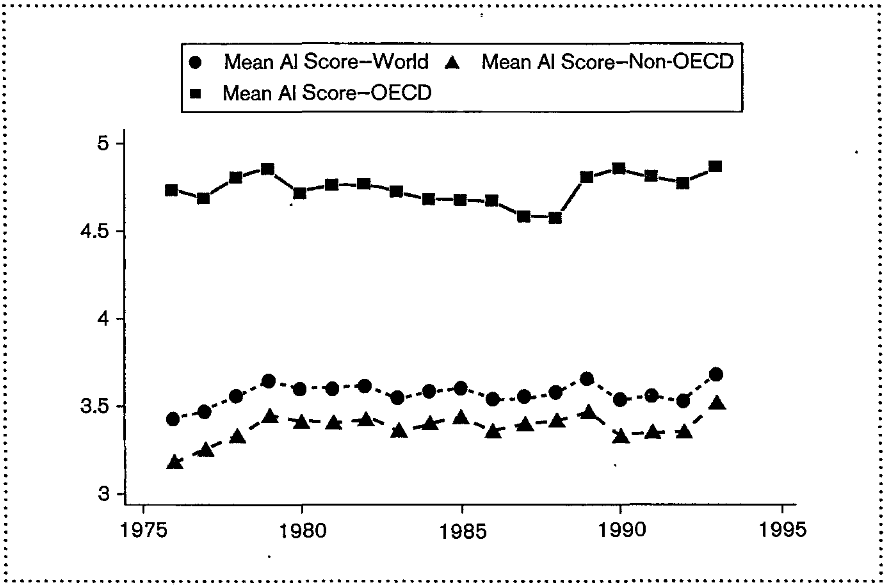
Milner, Poe, and Leblang plotted trends in rights using line graphs. Graphs are an important feature of presenting research findings (discussed in chapter 11). Trends in personal integrity rights are reported in figure 1-6; higher scores on the Amnesty International (AI) Human Rights Index indicate greater realization of rights. In the graph, data are presented for the world, OECD countries, and non-OECD countries. Notice that personal integrity rights worsened between 1989 and 1992 among non-OECD countries; this period corresponds with the outbreak of ethnic conflicts in Eastern Europe. The researchers found that globally, the trends in democratic rights, physical quality of life, and economic freedoms had been positive since 1975.

More recently, Minkler and Sweeney investigated whether developing countries respected security and subsistence rights simultaneously, as I/I would predict. To

[35] Wesley T. Milner, Stephen C. Poe, and David Leblang, "Security Rights, Subsistence Rights, and Liberties: A Theoretical Survey of the Empirical Landscape," *Human Rights Quarterly* 21, no. 2 (1999): 403–43.

[36] Lanse Minkler and Shawna Sweeney, "On the Indivisibility and Interdependence of Basic Rights in Developing Countries," *Human Rights Quarterly* 33 (2011): 351–96.

FIGURE 1-6 **Trends in Personal Integrity Rights**



Source: Wesley T. Milner, Stephen C. Poe, and David Leblang, "Security Rights, Subsistence Rights, and Liberties: A Theoretical Survey of the Empirical Landscape," *Human Rights Quarterly* 21, no. 2 (1999): 431, fig. 3. © 1999 The Johns Hopkins University Press. Reprinted with permission of The Johns Hopkins University Press.

do this, they used a statistical procedure called *correlation* (explained in chapter 13). Using a sample of 151 developing countries for the period 1997–2005, they found a modest, but statistically significant positive correlation between security and subsistence rights. They then set out to see if they could determine factors that accounted for the variation among developing nations—why security and subsistence rights were more closely connected in some developing countries than in others. Among the factors they considered were some they categorized as related to *ability* to advance basic rights (wealth, legal origins, and economic globalization) and others as related to *willingness* (democracy and government ratification of international human rights conventions), while taking into account variation in population size and involvement in internal and interstate conflicts. They found that a country's income, degree of trade openness, democratic political institutions, population size, and degree of internal conflict were all important factors in explaining why countries protected both types of basic rights simultaneously. A country's legal origins and endorsement of international human rights conventions played a lesser explanatory role, and degree of direct foreign investment and involvement in an international conflict played no role.

Our final example of research on human rights concerns the type of regime and the violation of civil liberties. Jørgen Møller and Svend-Erik Skaaning examine democracies and autocracies for differences in violation of freedom of expression, freedom of assembly and association, freedom of religion, and freedom of movement.[37] Not surprisingly, they find that democracies consistently score higher than autocracies in all four areas of civil liberties. What is surprising to the authors is that protection of civil liberties does not vary by autocratic subtypes: civilian, military, or monarchies (see figure 1-7). The authors argue that this finding suggests that there is no justification for supporting particular kinds of autocracies over others. Another interesting finding is that protection of civil liberties in democracies declined between 1997 and 2007, as shown in figure 1-8. The authors attribute this to the addition of new, less well-developed democracies to the group of democratic nations.

A Look into Judicial Decision Making and Its Effects

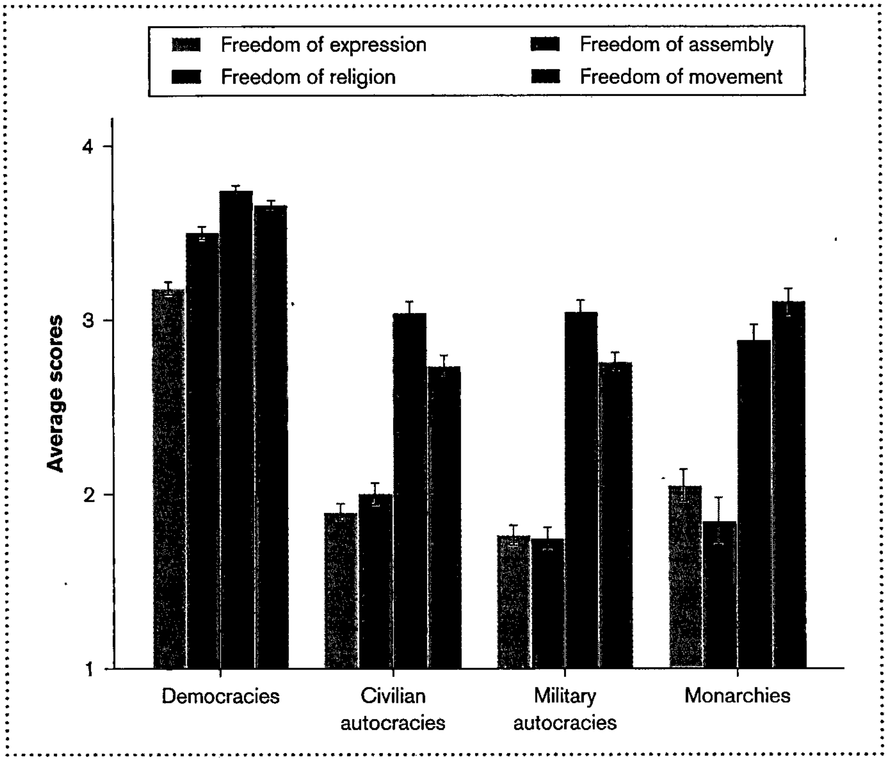
When the decisions of public officials clearly and visibly affect the lives of the populace, political scientists are interested in the process by which those decisions are reached. This is as true when the public officials are judges as when they are legislators or executives. One legal scholar stated, "given the often critical role judges play in our constitutional, political, and social lives, it is axiomatic that we need to better understand how and why judges reach the decisions they do in the course of discharging their judicial roles."[38] The decision-making behavior of the nine justices of the US Supreme Court is especially intriguing because they are not elected officials, their deliberations are secret, they serve for life, and their decisions constrain other judges and frequently have a major impact on national as well as state and local policy. No wonder political scientists try to determine just how Supreme Court justices reach their decisions.

The study of judicial decision making has been approached from several perspectives. Early studies investigated the influences of a judge's background (for example, as a prosecutor or defense attorney) and personal attributes such as race or gender. The results have been mixed, with little evidence to support the influence of these

[37] Jørgen Møller and Svend-Erik Skaaning, "Autocracies, Democracies, and the Violation of Civil Liberties," *Democratization* 20, no. 1 (2013): 82–106; DOI: 10.1080/13510347.2013.738863

[38] Michael Heise, "The Past, Present, and Future of Empirical Legal Scholarship: Judicial Decision Making and the New Empiricism," *University of Illinois Law Review* 2002, no. 4 (2002): 832. Available at http://illinoislawreview.org/wp-content/ilr-content/articles/2002/4/Heise.pdf

FIGURE 1-7 **Civil Liberty Protection by Regime Type**



Source: Jørgen Møller and Svend-Erik Skaaning, "Autocracies, Democracies, and the Violation of Civil Liberties," *Democratization* 20, no. 1 (2013): 82–106, figs. on p. 90; DOI: 10.1080/13510347.2013.738863

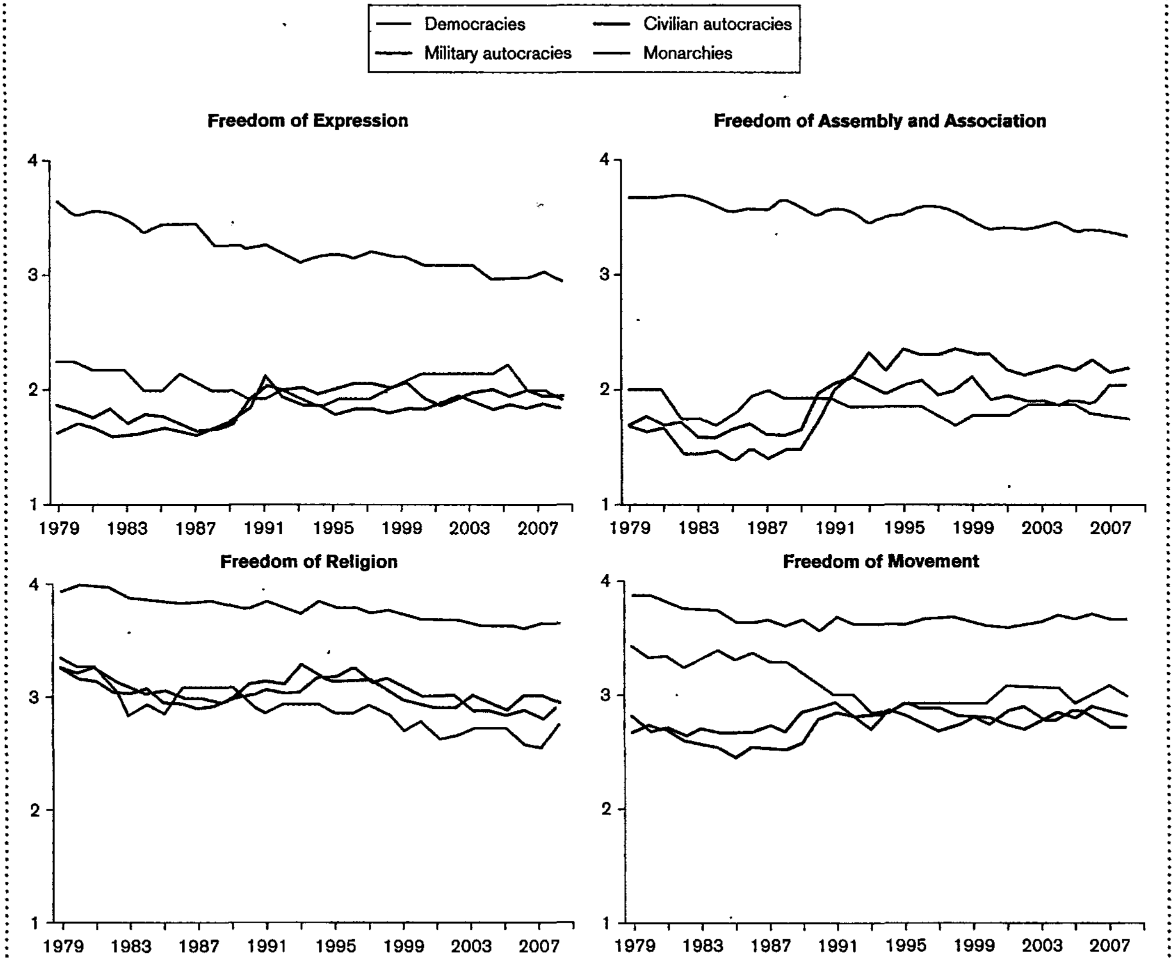
Note: 99 percent confidence interval.

factors.[39] One school of thought concerning judicial decision making holds that decisions are shaped primarily by legal doctrine and precedent. Because most Supreme Court judges have spent many years rendering judicial decisions while serving on lower courts, and because judges in general are thought to respect the decisions made by previous courts, this approach posits that the decisions of Supreme Court justices depend on a search for, and discovery of, relevant legal precedent.

Another view of judicial decision making proposes that judges, like other politicians, make decisions in part based on personal political beliefs and values. Furthermore, because Supreme Court judges are not elected, serve for life, seldom seek

[39] Ibid., 834–35.

FIGURE 1-8 **Trends in Civil Liberties Protection, 1979–2007**



Source: Jørgen Møller and Svend-Erik Skaaning, "Autocracies, Democracies, and the Violation of Civil Liberties," *Democratization* 20, no. 1 (2013): 82–106, figs. on pp. 91–94; DOI: 10.1080/13510347.2013.738863

any other office, and are not expected to justify their decisions to the public, they are in an ideal position to act in accord with their personal value systems.[40]

[40] For an example of research that considers both precedent and values, see Youngsik Lim, "An Empirical Analysis of Supreme Court Justices' Decision Making," *Journal of Legal Studies* 29, no. 2 (2000): 721–52.

One of the obstacles to discovering the relationship between the personal attitudes of justices and the decisions handed down by the Court is the difficulty of measuring judicial attitudes. Supreme Court justices do not often consent to give interviews to researchers while they are on the bench; nor do they fill out attitudinal surveys. Their deliberations are secret, they seldom make public speeches during their terms, and their written publications consist mainly of their case decisions. Consequently, about all we can observe of the political attitudes of Supreme Court justices during their terms are the written decisions they offer, which are precisely what researchers are seeking to explain. Some researchers use political party and the appointing president as indicators of judicial attitudes, although these are less than satisfactory measures.

An inventive attempt to overcome this obstacle is contained within Jeffrey A. Segal and Albert D. Cover's article "Ideological Values and the Votes of U.S. Supreme Court Justices."[41] Segal and Cover decided that an appropriate way to measure the attitudes of judges, independent of the decisions they make, would be to analyze the editorial columns written about them in four major US daily newspapers after their nomination by the president but before their confirmation by the Senate. This data source, the researchers argued, provides a comparable measure of attitudes for all justices studied, independent of the judicial decisions rendered and free of systematic errors. Here, too, though, the researchers had to accept a measure that was not ideal, for the editorial columns reflected journalists' perceptions of judicial attitudes rather than the attitudes themselves.

Despite this limitation, the editorial columns did provide an independent measure of the attitudes of the eighteen Supreme Court justices who served between 1953 and 1987. Segal and Cover found a strong relationship between the justices' decisions on cases dealing with civil liberties and the justices' personal attitudes as evinced in editorial columns. Those justices who were perceived to be liberal *before* their term on the Supreme Court voted in a manner consistent with this perception once they got on the Court. Judicial attitudes, then, do seem to be an important component of judicial decision making.

Other researchers have investigated the influence of so-called extralegal factors on the decisions of Supreme Court justices. Are there factors in addition to ideology but outside of legal precedent that influence judicial decision making? Do judges behave strategically to increase their prestige or influence vis-à-vis other judges and other branches of government?[42] Are they subject to influence by other judges and

[41] Jeffrey A. Segal and Albert D. Cover, "Ideological Values and the Votes of U.S. Supreme Court Justices," *American Political Science Review* 83, no. 2 (1989): 557–65.

[42] For an example of an investigation of strategic considerations, see Forrest Maltzman and Paul J. Wahlbeck, "Strategic Policy Considerations and Voting Fluidity on the Burger Court," *American Political Science Review* 90, no. 3 (1996): 581–92.

governmental actors? Among the possibilities are congressional influence (given the ability of Congress to pass legislation that overrides Court decisions and to initiate constitutional amendments, among other actions), presidential influence, and public opinion.[43]

The presidential election in 2000 brought into sharp relief for many Americans the importance of Supreme Court decisions to American politics. Some people felt that the high regard that Americans have for the Supreme Court brought closure to the highly contentious election and that support for the Supreme Court as an institution helped people to accept its decision in *Bush v. Gore* (2000). Others argued that general support and respect for the Supreme Court was undermined among those disappointed by the decision. Interestingly, political scientist Valerie J. Hoekstra was already busy investigating the two general questions raised so vividly by the 2000 decision: (1) How does the content of Supreme Court decisions affect support for the Court? That is, does respect for the Court decline among people who disagree with a decision? (2) Do Supreme Court decisions have any effect on public opinion? In other words, does the public change its mind about public policy issues once the Supreme Court has spoken?[44]

Hoekstra's work demonstrates how the choice of a research design (the topic of chapter 6) affects a researcher's ability to answer research questions with confidence. Hoekstra noted that public opinion polls generally show that the Supreme Court enjoys higher and more stable levels of public support than Congress or presidents, but that stability of aggregate-level measures such as public opinion polls does not mean that the opinions of individuals have not changed.[45] She argued that a panel study, one in which the same individuals are interviewed before and after a Supreme Court decision, is best to examine how support for the Supreme Court changes and whether individuals change their views about an issue in response to the Court's decision on a case. She also argued that it is important to interview individuals who are aware of the case to be decided by the Court. One cannot expect a decision of the Supreme Court to influence how people feel about an issue if people are not aware of the decision. Most Supreme Court decisions do not have the national significance and high level of public awareness as did *Bush v. Gore*. Therefore, Hoekstra selected four cases and interviewed people in the communities from which the cases originated.

[43] See Thomas G. Hansford and David F. Damore, "Congressional Preferences, Perceptions of Threat, and Supreme Court Decision Making," *American Politics Quarterly* 28, no. 4 (2000): 490–510; and Jeff Yates and Andrew Whitford, "Presidential Power and the United States Supreme Court," *Political Research Quarterly* 51, no. 2 (1998): 539–50.

[44] Valerie J. Hoekstra, *Public Reaction to Supreme Court Decisions* (New York: Cambridge University Press, 2003).

[45] Ibid., 13.

Hoekstra hypothesized that people who are more supportive of the Supreme Court are more likely to change their view of an issue in the direction of the Court's decision and that people who have strong opinions about an issue are less likely to change their views than are people whose opinions are not as strong. In two of the four cases, Hoekstra found that public opinion shifted in the direction of the Court's decision, but initial levels of support for the Court did not have an effect on the amount of change.[46] She did find that people who paid more attention to politics, and presumably were more aware of the issue, were more likely to change their opinion in the direction of the Court's decision.[47] Overall, she found limited support for the persuasive effect of Supreme Court decisions.

In terms of the effect of Supreme Court decisions on the public's support for the Court, Hoekstra found that people who were pleased with the Court's decision became more confident in and supportive of the Court, whereas those who were disappointed with the decision became less supportive. These changes were affected by how strongly a person felt about the issue: those who cared strongly about an issue tended to change their views of the Court more than those who did not care as much about the issue.[48]

Because of changes in the membership of the Supreme Court, current researchers are able to examine whether the persuasiveness of the Supreme Court is mediated by partisanship. As Lawrence Baum and Neal Devins point out, "for the first time in more than a century, the ideological positions of the justices on today's Supreme Court can be identified purely by party affiliation."[49] Consequently, political scientists are able to investigate the role of partisanship in support for Supreme Court decisions and whether the public perceives the Supreme Court as a partisan or political decision-making institution rather than a legalistic, neutral, or impartial one.

Stephen P. Nicholson and Thomas G. Hanford used survey research and experimentation to investigate these topics.[50] They wanted to find out if public acceptance of four relatively recent Supreme Court decisions was influenced by cues they imbedded in survey questions. For example, when respondents were asked if they agreed with a decision, some versions of the question just mentioned that the

[46] Ibid., 113.

[47] Ibid., 114.

[48] Ibid., 137.

[49] Lawrence Baum and Neal Devins, "Split Definitive: For the First Time in a Century, the Supreme Court Is Divided Solely by Political Party," *Slate*. Accessed January 13, 2015. Available at http://www.slate.com/articles/news_and_politics/jurisprudence/2011/11/supreme_court_s_partisan_divide_and_obama_s_health_care_law.html

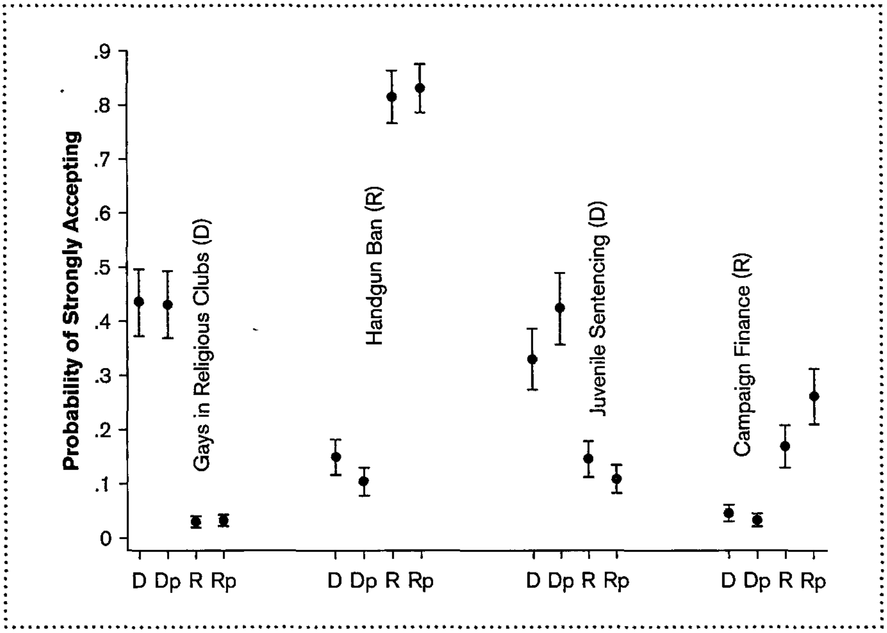
[50] Stephen P. Nicholson and Thomas G. Hansford, "Partisans in Robes: Party Cues and Public Acceptance of Supreme Court Decisions," *American Journal of Political Science* 58, no. 3 (2014): 620–36.

decision had been made by "the government," while other versions of the question attributed the decision to the Supreme Court or to a Republican-appointed (or Democratic-appointed) majority on the Supreme Court. They chose to ask respondents about four different decisions: *Christian Legal Society v. Martinez* (2010), in which the Court determined that a law school may require that religious student clubs admit gay students; *District of Columbia v. Heller* (2008), in which the Court struck down Washington, D.C.'s handgun ban; *Graham v. Florida* (2010), in which the Court ruled that juveniles cannot be sentenced to life without parole for any crime other than murder; and *Citizens United v. FEC* (2010), in which the Court held that independent campaign expenditures cannot be limited. Two of these decisions involved high levels of partisan polarization (gun control and gays in religious clubs), and two exhibited low levels (limits on the sentencing of juveniles and campaign finance). For two of the issues, they found a statistically significant, but small increase in support when the decision was attributed to the Supreme Court as opposed to the government. For the other two, there was no difference, which suggests limited support for the idea that the public gives the Supreme Court greater deference than other government decision-making bodies. Figure 1-9 presents the results comparing support for the decisions when the party of the majority of the justices deciding is revealed and when it is not for respondents who are either strong Democrats or strong Republicans. It is clear from the figure that Republicans and Democrats differ in their support for Court decisions, with the partisan divide greatest for gays in religious clubs and the handgun ban. Adding the cue of the partisanship of the majority of the Court did not change the acceptance level among respondents for this issue, but notice that when the partisan cue is added for the three other decisions, the gap between Republicans and Democrats increases. Overall, Nicholson and Hansford conclude that their findings indicate that party cues have a greater effect on acceptance of Supreme Court decisions when the issues are not highly polarized and that the public "perceives the contemporary Supreme Court as similar to other partisan actors, at least with regard to public acceptance of its decisions." For two of the four policy outcomes in their experiment, attribution of the policy decision to the Supreme Court, rather than to the government in general, increases acceptance of the decision, but only slightly. This is scant support for the idea that the public views the Court as a legal institution and raises questions over compliance with Supreme Court decisions, which relies on public acceptance and perceptions of the Court as an apolitical, nonpartisan institution.

Influencing Bureaucracies

Disasters such as the deaths of twenty-nine miners in an explosion in the Upper Big Branch mine in Montcoal, West Virginia, on April 5, 2010, and the explosion of BP's Deepwater Horizon oil rig on April 20, 2010, which killed eleven workers,

FIGURE 1-9 **Probability of Strong Acceptance by Party ID and Party Cue**



Source: Stephen P. Nicholson and Thomas G. Hansford, "Partisans in Robes: Party Cues and Public Acceptance of Supreme Court Decisions," *American Journal of Political Science* 58, no. 3 (2014): fig. 2, pp. 620–36.

Note: *D* and *R* on the *x*-axis represent strong Democrat or strong Republican subject, and *p* indicates the presence of the *Party Cue*. *D* and *R* in the decision label reveal the partisan identity of the majority behind the decision. For all the predicted probabilities presented here, *Supreme Court* is held at one. Bars are 95 percent confidence intervals.

injured seventeen, and spilled an estimated 4.9 million barrels of crude oil into the Gulf of Mexico, tragically brought the performance of two federal bureaucracies into the limelight. In the case of the mining disaster, questions were raised about whether the US Mine Safety and Health Administration effectively enforced mining regulations. In the case of the Deepwater Horizon explosion, the US Minerals Management Service was being criticized for being too closely aligned with the oil industry and, among other things, insufficient scrutiny of oil spill response plans.

Political control of bureaucracy is an ongoing topic of discussion and investigation by political scientists. A variety of theories and beliefs about political influence on bureaucratic activities have ascended, only to be superseded by new theories and beliefs based on yet more research. Theories have evolved from the politics versus administration dichotomy, which strictly separates politics and administration and argues that the way to avoid problems such as political patronage and corruption in administration is to pursue professionalism and independence in administration, to the iron triangle (or capture) theory, which argues that administration and politics are inseparable and views agencies as responsive to a narrow range of advantaged and special interests assisted by a few strategically located members of Congress. This theory raises serious questions about democratic control of government agencies. A more recent theory, principal-agent theory, suggests that presidents and Congress (the principals) do have ways to control bureaucratic activities or agents. According to this theory, policy makers use rewards or sanctions to bring agency activities back in line when they stray too far from the policy preferences of elected politicians. Control mechanisms include budgeting, political appointments, structure and reorganization, personnel power, and oversight.[51] Research shows that agency outputs vary with political changes. The emergence of a new presidential administration, the seating of new personnel on the courts, and change in the ideological stances of congressional oversight committees all influence agency outputs.[52] Research also indicates that presidents and Congress compete over the control of agencies and that agencies vary in the extent to which they are designed to be insulated from presidential control.[53] Other research, however, presents evidence that bureaucratic values may be more influential than political control mechanisms.[54]

Richard L. Hall and Kristina C. Miler noted that Congress tries to compensate for the difficulty of overseeing agency decisions by designing procedural requirements to improve the visibility of decision making, openness to multiple points of view, and accountability (by including standing to sue, for example). Sometimes, Congress will limit agency discretion specifically through statutes.[55] Yet questions remain about who influences agency decisions and by what means. Hall and Miler investigated the circumstances surrounding the decisions by members of Congress to intervene in agency decisions ex poste—that is, to decide whether to challenge, or defend in the face of a challenge, a particular agency rule—and the role that interest groups play in those decisions.

[51] This discussion is based on B. Dan Wood and Richard W. Waterman, "The Dynamics of Political Control of the Bureaucracy," *American Political Science Review* 85, no. 3 (1991): 801–28.

[52] Ibid.

[53] David E. Lewis, *Presidents and the Politics of Agency Design: Political Insulation in the United States Government Bureaucracy, 1946–1997* (Stanford, Calif.: Stanford University Press, 2003).

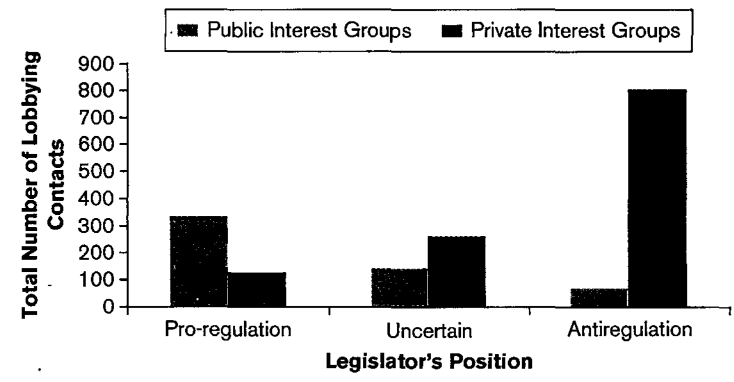
[54] See Kenneth J. Meier and Laurence J. O'Toole Jr., "Political Control versus Bureaucratic Values: Reframing the Debate," *Public Administration Review* 66, no. 2 (2006): 178–92; and Martha Wagner Weinberg, *Managing the State* (Cambridge, Mass.: MIT Press, 1977).

[55] Richard C. Hall and Kristina Miler, "What Happens after the Alarm? Interest Group Subsidies to Legislative Overseers," *Journal of Politics* 70, no. 4 (2008): 990–1005.

Oversight behaviors include writing a letter, submitting comments, giving a speech, introducing a bill, offering an appropriations rider, and challenging (or defending) an agency policy during a congressional oversight hearing. Such behaviors are costly to legislators; gathering issue-specific information consumes time and labor that could be spent on other priorities, and, as Hall and Miler explained, constituents may not notice and reward oversight activity. Oversight activity is easier for legislators serving on the relevant committees and subcommittees because they have more staff and issue-specific expertise.

Hall and Miler hypothesized that lobbyists "subsidize" legislative oversight by providing labor and information to legislators. Pursuing the subsidy perspective, they also hypothesized that lobbyists will target legislative allies rather than try to persuade uncertain legislators unfriendly to their position. To test their hypotheses, they used the case of the 1997 fight over the Environmental Protection Agency's proposal to strengthen air-quality standards for ground-level ozone and particulate matter. They interviewed six of eight principal lobbyists on the pro-regulation side and nine of fourteen on the antiregulation side. They asked lobbyists how many times they had contacted each member on the House Commerce Committee, which had oversight jurisdiction over the EPA. The number of contacts made by industry lobbyists was about twice the number of contacts made by health and environmental coalition (HEC) lobbyists. The pattern of contacts is shown in figure 1-10. As expected according to the subsidy hypothesis, lobbyists focused their contacts on "friendly" legislators—those who were allied with their side of the issue.

FIGURE 1-10 **Lobbying the House Commerce Committee: Public Interest Groups, Private Interest Groups, and the 1997 NAAQS Regulations**

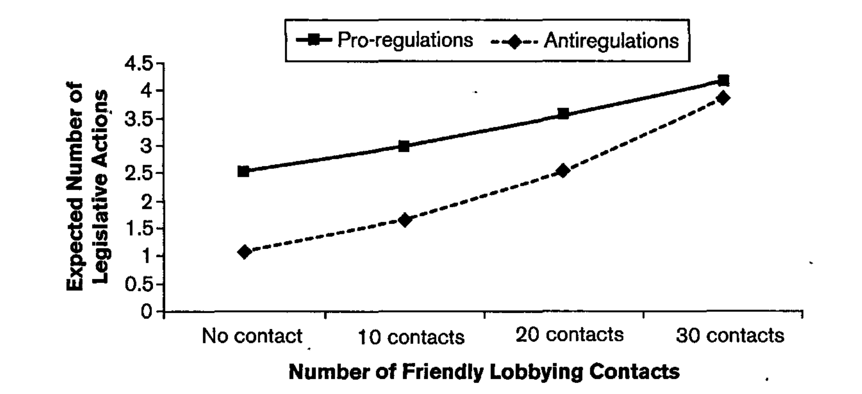


Source: Richard L. Hall and Kristina C. Miler, "What Happens after the Alarm? Interest Group Subsidies to Legislative Overseers," *Journal of Politics* 70, no. 4 (2008): fig. 1, p. 997. Reproduced with the permission of Cambridge University Press.

The researchers also found that the more HEC groups lobbied pro-environment legislators, the more likely those legislators were to send comments to the EPA in support of the stricter regulations. Similarly, the more industry groups lobbied pro-industry legislators, the more likely those legislators were to send comments to the EPA critical of those regulations. Figure 1-11 shows the impact of lobbying contact on the number of comments made by friendly committee members. So, for example, a legislator who was contacted thirty times by industry lobbyists would make about 4 antiregulation comments, while a legislator who received no contact from industry lobbyists would make 2.5 comments. HEC-friendly legislators who received thirty contacts from HEC lobbyists made almost 4 comments, while those who received no contacts made an average of 1 comment.

Hall and Miler also investigated the possibility that campaign contributions played a role in the commenting behavior of committee members and found that they had little or no impact. The researchers concluded that lobbying targeted at friendly legislators was effective for both environmental and industry lobbyists. Lobbying targeted at friendly legislators who are subcommittee and committee leaders was especially effective. Industry lobbyists were at an advantage because they were able to contact legislators more (they contacted nine members more than thirty times, and some of them eighty times, while HEC lobbyists contacted only one member more than thirty times) and because industry-friendly legislators were full committee leaders. While this research looked at lobbying and oversight in only one case, it is a case involving highly technical issues that are difficult or costly for legislators to oversee and, therefore, a good one to use to test the subsidy theory of lobbying.

FIGURE 1-11 **Effects of Friendly Lobbying on Legislative Activity**



Source: Richard L. Hall and Kristina C. Miler, "What Happens after the Alarm? Interest Group Subsidies to Legislative Overseers," *Journal of Politics* 70, no. 4 (2008): fig. 2, p. 1002. Reproduced with the permission of Cambridge University Press.

Effects of Campaign Advertising on Voters

Enormous sums of money are spent on campaign advertising by candidates vying for political office. Political scientists have long been interested in the effects of campaign advertising on voters. Some have argued that advertising has little effect, due to the public's ability to screen out messages conflicting with their existing views. Others have suggested that campaign activity, including advertising, stimulates voter interest and increases turnout. Still others suggest that negative campaign advertising, particularly television advertisements, has harmful effects on the democratic process: negative campaign ads are thought to increase cynicism about politics and to cause the electorate to turn away from elections in disgust, a phenomenon called *demobilization*. Since the Supreme Court's 2010 decision in *Citizens United v. Federal Election Commission*, which removed many restrictions on campaign advertising by outside or independent groups, researchers have been interested in the impact of independent spending on negative campaign ads. Let's start with a look at some of the earlier research, then examine some of the latest research.

A 1994 study on so-called attack advertising by Stephen D. Ansolabehere, Shanto Iyengar, Adam Simon, and Nicholas Valentino is widely recognized as establishing support for the demobilization theory. Noting that "more often than not, candidates criticize, discredit, or belittle their opponents rather than providing their own ideas," the researchers hypothesized that, rather than stimulating voter turnout, such campaigns would depress turnout.[56]

Ansolabehere and his colleagues devised a controlled experiment in which groups of prospective voters were exposed to one of three advertisement treatments: positive political advertisements, no political advertisements, or negative political advertisements. After taking into account other factors likely to affect a person's intention to vote, the researchers found that exposure to negative (as compared to positive) advertisements depressed intention to vote by 5 percent.

Recognizing that the size of the experimental effect—that is, how much impact advertising has on behavior—might not match the size of the real-world effect, the researchers also devised a strategy to measure the effect of negative advertising in real campaigns. They measured the tone of the campaigns in the thirty-four states that held a Senate election in 1992. They calculated the turnout rate and something called the "roll-off rate" for each Senate race. The roll-off rate measures the extent to which people who were sufficiently motivated to vote in the presidential election chose not to vote in the Senate race. The researchers found that both the turnout rate and the roll-off rate were affected by campaign tone. Turnout in states with

[56] Stephen D. Ansolabehere, Shanto Iyengar, Adam Simon, and Nicholas Valentino, "Does Attack Advertising Demobilize the Electorate?" *American Political Science Review* 88, no. 4 (1994): 829–38. Available at http://weber.ucsd.edu/~tkousser/Ansolabehere.pdf

a positive campaign tone was 4 percent higher than in states where the tone was negative. The difference in roll-off rates was 2.4 percent, with roll-off rates higher in those states with more negative campaign advertising. These results confirmed the team's earlier results and demonstrated that negative campaigns may in fact depress voter turnout.

Ansolabehere and his colleagues suggested that the decline in presidential and midterm voter turnout since 1960 may be due in part to the increasingly negative tone of national campaigns. They also raised some interesting questions, asking whether or not candidates should "be free to use advertising techniques that have the effect of reducing voter turnout" and whether or not "in the case of publicly financed presidential campaigns, [it is] legitimate for candidates to use public funds in ways that are likely to discourage voting."[57]

Subsequent researchers have conducted studies using different approaches that qualify this finding. For example, Martin P. Wattenberg and Craig Leonard Brians investigated the contention that "the intent of most negative commercials is to convert votes by focusing on an issue for which the sponsoring candidate has credibility in handling but on which the opponent is weak."[58] Using survey or poll data from the 1992 and 1996 US presidential elections that allowed the identification of respondents who recalled seeing negative ads, positive ads, or no ads at all and the comparison of their turnout rates, Wattenberg and Brians found that negative ads did not depress turnout. In fact, for groups considered unlikely to vote (such as young people or those lacking a high school education), turnout rates were higher for those who recalled seeing either a positive or negative ad, compared to those who recalled no ad. For groups expected to have higher turnout rates, ad recall had only a slight effect on turnout rates. After taking into account a wide range of factors associated with turnout, the researchers found that recall of negative political ads was significantly associated with higher turnout rates in the 1992 elections. For the 1996 elections, they found that recall of ads, whether positive or negative, had no impact on turnout rates. They also concluded that recalling a negative ad did not have a depressing effect on a person's sense of political efficacy.

They suggested that the experimental findings of Ansolabehere and his colleagues do not hold up in the real world of elections. Recall, though, that those experimental findings were buttressed by the analysis of aggregate voting data in the 1992 Senate races. Wattenberg and Brians questioned these findings and pointed out that the election data used by Ansolabehere and his colleagues are different from the official 1992 election returns published by the Federal Election Commission (FEC).

[57] Ibid., 835.

[58] Martin P. Wattenberg and Craig Leonard Brians, "Negative Campaign Advertising: Demobilizer or Mobilizer?" *American Political Science Review* 93, no. 4 (1999): 891. Available at http://weber.ucsd.edu/~tkousser/Wattenberg.pdf

As we noted earlier in the chapter, political science is an iterative or cumulative activity and often involves debates over measurement of variables. Ansolabehere and his colleagues responded to Wattenberg and Brians's study by noting that survey recall data are prone to inaccuracies: recall is a poor measure of actual exposure, and people who are likely to vote are more likely to recall seeing a political ad.[59] They analyzed the survey data for the 1992 and 1996 elections, making adjustments for exposure to campaign ads that Wattenberg and Brians did not. They used data measuring the volume of ads in the different senatorial elections, noting that higher-volume campaigns have disproportionately more negative ads. They also noted that the tone of campaigns becomes more negative as elections approach. Thus, respondents surveyed earlier in an election will have been exposed to less negative campaigning than those interviewed later in an election. Their analysis showed that recall of negative ads was significantly higher in states with higher levels of advertising and in the latter stages of the campaign, and that intention to vote was lower in states with more television advertising and in the latter stages of campaigns. They therefore concluded that negative advertising has a negative impact on voter turnout. They also replicated their analyses of the Senate races using official FEC data (previously they had used data obtained directly from the election officers in each state) and concluded that, on average, turnout in positive campaigns is nearly 5 percentage points higher than turnout in negative campaigns.

In 2009 Richard R. Lau and Ivy Brown Rovner commented on the "explosion" of research on negative campaigning over the previous two decades.[60] They reported that by the end of 2006, there were 110 books, chapters, dissertations, and articles on the effects of negative political advertisements or negative campaigns, and many more exploring other aspects of negative campaigns. One might think that, in the twenty-plus years since the early research investigating the demobilization hypothesis, the topic would have run its course. Quite the contrary. As they report, there are many remaining challenges. Among the difficulties researchers face are reconciling their definitions of negative advertising with what the public considers to be fair or unfair attack ads; measuring negativity based only on the text of an ad rather than its accompanying visuals and music, which also contribute to the tone of an ad; measuring the impact of negative campaigning on the outcome of races when campaigns are most likely to go negative in closely contested races; measuring the tone of all campaign-related material (television ads, phone calls, campaign mail, and personal contacts); taking into account both tone (the proportion of negative ads) as well as volume (the total number of ads shown in a media market); determining the effect of negative ads on the outcome of a single race or type of

[59] Stephen D. Ansolabehere, Shanto Iyengar, and Adam Simon, "Replicating Experiments Using Aggregate and Survey Data: The Case of Negative Advertising and Turnout," *American Political Science Review* 93, no. 4 (1999): 901–10.

[60] Richard R. Lau and Ivy Brown Rovner, "Negative Campaigning," *Annual Review of Political Science* 12 (2009): 285–306.

race when in most cases multiple election contests are occurring simultaneously; and determining the effect of negative campaign ads on the political system as a whole when studying the amount of negative campaigning in a single race or type of race. In addition to the need for more research to address these issues, much has changed in the world of negative campaign ads since Lau and Rovner wrote their assessment.

TABLE 1-1 **Independent Expenditures, 2000–2014**

Year	Independent Expenditures
2014	549,905,830
2012	1,002,135,419
2010	205,519,230
2008	143,618,022
2006	37,801,719
2004	63,885,795
2002	16,747,650
2000	33,778,636

Source: Data from the Center of Responsive Politics. Accessed June 7, 2015. Available at https://www.opensecrets.org/outsidespending/cycle_tots.php

More recently, research on negative campaign advertising has focused on the effects of the dramatic increase in spending by independent groups following the 2010 Supreme Court decision in *Citizens United v. Federal Election Commission,* particularly by independent- expenditure-only committees, commonly known as "super PACS." According to the Center for Responsive Politics, spending on independent expenditures, ads that expressly advocate the election or defeat of specific candidates and are aimed at the electorate as a whole, topped one billion dollars in 2012 and were over half a billion in 2014 (a nonpresidential election year; see table 1-1). Much of this increase in spending has been on negative ads.

Not only has spending of negative ads increased dramatically, but the ability of voters to determine who is behind the ads is limited. Groups sponsoring the ads use deliberately vague and appealing names, and the identity of donors is not revealed. The public and researchers alike wonder about the impact of the increase in spending on negative ads. In particular, some researchers want to know if revealing ad sponsors and the identity of donors changes the impact of negative ads and whether the way in which information about the donors is revealed makes a difference.

Deborah Jordan Brooks and Michael Murov investigated how the public responds to ads sponsored by candidates as compared to ads sponsored by super PACS and other independent groups.[61] Specifically, they examined whether harsh ads were more effective if they were sponsored by independent groups than if they were sponsored by candidates. Previous research indicates that there is a backlash effect against candidates for running negative ads against their opponents. It is also possible that the public finds negative ads sponsored by candidates more persuasive because candidate-sponsored ads must include a statement by the candidate endorsing the ad. The fact that candidates have to associate themselves with an ad may make the claims

[61] Deborah Jordan Brooks and Michael Murov, "Assessing Accountability in a Post-*Citizens United* Era: The Effects of Attack Ad Sponsorship by Unknown Independent Groups," *American Politics Research* 40, no. 3 (2012): 383–418.

more credible. Brooks and Murov designed an experiment in which viewers were shown ads in which the only factor that varied was the ad attribution. First, subjects were shown two positive ads about the candidates in a fictional state assembly race and asked to evaluate the candidates. Then they were shown an ad attacking the personal traits of a candidate and again asked to evaluate the candidates. They found that subjects reacted differently depending on the sponsorship of the ad, and that the differences were due to backlash, not persuasion. The trait-based attack ad sponsored by an unknown independent group was more effective than the one sponsored by a candidate because candidates were "punished" for running attack ads.

Although in *Citizens United* the Supreme Court upheld disclosure requirements, current law does not require all independent groups to disclose their donors, and so far Congress has failed to pass legislation closing the loopholes. So if independent groups are less likely than candidates to be held accountable for their attack ads, well-funded groups not reporting their donors can run attacks (truthful or not) and have an important impact on elections. The results of the study by Brooks and Murov raise the questions, "How can independent groups be held accountable for their ads?" and "How can viewers be given information that helps them discern the interests and motivations of ad sponsors?" These are some of the very questions that are investigated by two other political scientists, Conor M. Dowling and Amber Wichowsky. They found that participants in their study were more supportive of the attacked candidate if they were informed that donors were anonymous or were given information about the donors, suggesting that voters may discount a group-sponsored ad when they have information about the financial interests behind the ad.[62] But Dowling and Wichowsky also found that the manner in which the information about donors was presented mattered. Their research, therefore, has very practical significance, as it can be used by policy makers to structure new disclosure requirements.

In chapter 6 we discuss some ways to design research to investigate the effects of advertising on political behavior. We simply note for now that this issue will surely continue to preoccupy researchers and illustrates some of the complexities and excitement of the empirical study of politics.

Research on Public Support for US Foreign Involvement

The ongoing wars in Iraq and Afghanistan highlight the relevance of research into public support for US military involvement in foreign affairs. Researchers have

[62] Conor M. Dowling and Amber Wichowsky, "Does It Matter Who's Behind the Curtain? Anonymity in Political Advertising and the Effects of Campaign Finance Disclosure," *American Politics Research* 41, no. 6 (2013): 965-96.

investigated a wide range of factors associated with public support for US military involvement such as attributes of individuals, including attitudes toward the use of military force and US involvement in world affairs in general, education, and knowledge of foreign affairs. Others factors include situational factors, such as the primary purpose or objective of US military involvement, the relative power of the United States vis-à-vis an adversary, the costs of involvement (particularly US military casualties), the extent of elite consensus over whether the United States should be involved, and multilateral support for involvement.[63] Let's take a look at one particularly relevant example that investigates the public's willingness to expend additional resources, both human and financial, in an ongoing war.

In an article titled "'Don't Let Them Die in Vain': Casualty Frames and Public Tolerance for Escalating Commitment in Iraq," William A. Boettcher III and Michael D. Cobb investigated the extent to which the public responds to rhetoric to the effect that the "sunk costs" or "sacrifices" made by the men and women killed in war must be redeemed through further conflict.[64] They pointed out that while there may be logical and rational reasons for continuing the fighting in Iraq, spent money and dead soldiers cannot be recovered by additional deaths and more spending. They noted that such a rhetorical argument (which they call "investment framing") appeals to a well-known and researched psychological bias called the "sunk-cost trap" in which "individuals pursue irrational and costly courses of action to redeem losses that cannot be recovered" or "good money is thrown after bad."[65]

They set out to test whether or not "investment frames" increase the public's willingness to continue the war (defined as willingness to tolerate additional casualties and to spend more money) and whether or not it makes a difference if well-known figures with credibility make the argument. They also took into consideration if individuals felt the United States "did the right thing" by going to war or if they felt the United States "should have stayed out of Iraq." Thus, they hypothesized that "the casualty and spending tolerance of individuals supportive of the decision to go

[63] For example, see Bruce Jentleson, "The Pretty Prudent Public: Post-Vietnam American Opinion on the Use of Military Force," *International Studies Quarterly* 36, no. 1 (1992): 49–74; Eric Larson, *Casualties and Consensus: The Historical Role of Casualties in Domestic Support for U.S. Military Operation* (Santa Monica, Calif.: RAND, 1996); Steven Kull, I. M. Destler, and Clay Ramsay, *The Foreign Policy Gap: How Policymakers Misread the Public* (Washington, D.C.: Center for Strategic and International Studies, 1997); Miroslav Nincic, "Domestic Costs, the U.S. Public, and the Isolationist Calculus," *International Studies Quarterly* 41, no. 4 (1997): 593–610; Richard K. Herrmann, Philip E. Tetlock, and Penny S. Visser, "Mass Public Decisions to Go to War: A Cognitive-Interactionist Framework," *American Political Science Review* 93, no. 3 (1999): 553–74; and Bruce W. Jentleson and Rebecca L. Britton, "Still Pretty Prudent: Post-Cold War American Public Opinion on the Use of Military Force," *Journal of Conflict Resolution* 42, no. 4 (1998): 395–417.

[64] William A. Boettcher and Michael D. Cobb, "'Don't Let Them Die in Vain': Casualty Frames and Public Tolerance for Escalating Commitment in Iraq," *Journal of Conflict Resolution* 53, no. 5 (2009): 677–97.

[65] Ibid., 678.

to war in Iraq will increase when exposed to investment frames, while the casualty and spending tolerance of individuals opposed to the decision to go to war in Iraq will be unaffected or decrease when exposed to investment frames."[66]

In a survey of 1,342 individuals of a representative sample of US households, respondents were given a battery of questions about Iraq. Then they were assigned to a control group or to one of several experimental groups in which the "investment framing" conditions were varied. In the unattributed investment frame condition, respondents read, "Some people say we need to stay and complete the mission in Iraq to honor the dead and make sure they did not die in vain." In two other conditions, the phrase "some people" was replaced with either "General Casey, the Commanding General in Iraq," or "Pat Robertson, founder of the Christian Coalition." A fourth, an alternative "consumer" frame, was attributed to Pope Benedict and discouraged respondents from honoring sunk costs by saying, "Staying will not bring them back and will only result in more loss of life."

While we cannot explore all of Boettcher and Cobb's results, they found that the investment frames had a positive impact on tolerance for additional casualties and spending among those who supported going to war in Iraq, with the unattributed frame having the most consistent impact. The investment frame attributed to General Casey had an especially negative impact on the tolerance of those opposed to going to war in the first place. The researchers surmised that respondents were discounting the investment argument because coming from this source, it was perceived as self-serving. Overall, they concluded that investment frames are counterproductive unless targeted at sympathetic audiences.

Douglas L. Kriner and Francis X. Shen also used an experiment imbedded in a survey to see if Americans are sensitive not only to the number of combat casualties in war but also to the distribution of those casualties across society.[67] They argued that Americans are likely to see inequalities in the distribution of war casualties as unfair because it violates their belief in political equality. In the experiment, respondents in a telephone poll were told the number of American casualties in the Korean, Vietnam, and Iraqi wars. Those assigned to the control group received no further information. A second group was given the "inequality" treatment, in which they were told that poorer communities suffered higher casualty rates in those wars than had wealthy communities. A third group was told that casualties had been shared by all communities. Then respondents in all groups were presented with four hypothetical military missions for which they were asked to provide a number of acceptable casualties. Kriner and Shen found that "levels of casualty sensitivity varied significantly across mission types and experimental treatments." In three of

[66] Ibid., 683.

[67] Douglas L. Kriner and Francis X. Shen, "Reassessing American Casualty Sensitivity: The Mediating Influence of Inequality," *Journal of Conflict Resolution* 58, no. 7 (2014): 1174–1201.

TABLE 1-2 Casualty Sensitivity by Mission Type and Inequality Cue

	Control (%)	Inequality (%)	Shared Sacrifice (%)
Liberia (internal policy change)			
High casualty sensitivity (0–50 casualties)	40.2	50.0	43.6
Moderate casualty sensitivity (51–5,000 casualties)	45.7	36.6	39.2
Low casualty sensitivity (>5,000 casualties)	14.1	13.4	17.2
Darfur (humanitarian intervention)			
High casualty sensitivity (0–50 casualties)	33.9	34.8	36.3
Moderate casualty sensitivity (51–5,000 casualties)	52.1	51.4	51.8
Low casualty sensitivity (>5,000 casualties)	14.0	13.8	11.9
Iran (foreign policy restraint)			
High casualty sensitivity (0–50 casualties)	28.7	*37.1*	*26.3*
Moderate casualty sensitivity (51–5,000 casualties)	49.8	43.2	51.0
Low casualty sensitivity (>5,000 casualties)	21.5	19.6	22.7
Al Qaeda in Somalia (foreign policy restraint/war on terror)			
High casualty sensitivity (0–50 casualties)	21.8	**32.6**	26.2
Moderate casualty sensitivity (51–5,000 casualties)	55.9	53.9	52.9
Low casualty sensitivity (>5,000 casualties)	22.3	**13.5**	21.0

Source: Douglas L. Kriner and Francis X. Shen, "Reassessing American Casualty Sensitivity: The Mediating Influence of Inequality," *Journal of Conflict Resolution* 58, no, 7 (2014): tab. 1, pp. 1174–201.

Notes: All percentages constructed using survey weights. Columns may not sum to 100 percent because of rounding. Percentages in boldface are significantly different from the control, p < .10. Percentages in italics are significantly different from the other treatment group, p < .10.

the four scenarios cues about inequalities in sacrifice significantly influenced casualty sensitivity, as shown in table 1-2.

In addition, Kriner and Shen were able to identify where respondents lived and thus divide them into two groups— those from states in the top half of the state casualty rate in the Iraq war and those in the bottom half. They found that respondents in the high-casualty-rate states were more sensitive to the inequality of treatment than were those in lower-casualty-rate states. The authors also report results of a 2011 poll in which 1,009 Americans were asked what parts of the United States they thought

American soldiers who have died fighting in Afghanistan and Iraq came from. Forty-five percent chose the correct response: "More casualties are coming from poor, less educated parts of the country"; 3 percent chose "from rich, more educated" areas; while 44 percent thought there was no difference in sacrifice between rich and poor communities.[68] Given the results of their earlier research, one has to wonder what would be the impact on US foreign policy if the American public was fully aware of the inequality in the distribution of war casualties.

Clearly, both citizens and politicians have quite a bit to learn from recent political science research on the conditions under which the public will support the use of military force and foreign policies advocated by national political leaders. It is exciting for researchers to investigate these issues and to pursue greater understanding of these and related questions.

Conclusion

Political scientists are continually adding to and revising our understanding of politics and government. As the several examples in this chapter illustrate, empirical research in political science is useful for satisfying intellectual curiosity and for evaluating real-world political conditions. New ways of designing investigations, the availability of new types of data, and new statistical techniques contribute to the ever-changing body of political science knowledge. Conducting empirical research is not a simple process, however. The information a researcher chooses to use, the method that he or she follows to investigate a research question, and the statistics used to report research findings may affect the conclusions drawn. For instance, some of these examples used sample surveys to measure important phenomena such as public opinion on a variety of public policy issues. Yet surveys are not always an accurate reflection of people's beliefs and attitudes. In addition, how a researcher measures the phenomena of interest can affect the conclusions reached. Finally, some researchers conducted experiments in which they were able to control the application of the experimental or test factor, whereas others compared naturally occurring cases in which the factors of interest varied.

Sometimes, researchers are unable to measure political phenomena themselves and have to rely on information collected by others, particularly government agencies. Can we always find readily available data to investigate a topic? If not, do we choose a different topic or collect our own data? How do we collect data firsthand? When we are trying to measure cause and effect in the real world of politics, rather than in a carefully controlled laboratory setting, how can we be sure that we have identified all the factors that could affect the phenomena we are trying to explain?

[68] Ibid., 1178.

Finally, do research findings based on the study of particular people, agencies, courts, communities, or countries have general applications to all people, agencies, courts, communities, or countries? To develop answers to these questions, we need to understand the process of scientific research, the subject of this book.

TERMS INTRODUCED

Applied research. Research designed to produce knowledge useful in altering a real-world condition or situation.

Empirical research. Research based on actual, "objective" observation of phenomena.

Pure, theoretical, or recreational research. Research designed to satisfy one's intellectual curiosity about some phenomenon.

CHAPTER 2

The Empirical Approach to Political Science

CHAPTER OBJECTIVES

2.1 Identify eight characteristics of empiricism.

2.2 Discuss the importance of theory in empiricism.

2.3 Explain the five steps in the empirical research process.

2.4 Describe practical obstacles that challenge the empirical approach.

2.5 Summarize competing perspectives.

POLITICAL SCIENTISTS JEFFREY WINTERS AND BENJAMIN Page wonder if the United States, despite being a nominal democracy, is not in fact governed by an oligarch, a relatively small number of very wealthy individuals and families.[1] Their work leads them to conclude:

> We believe it is now appropriate to . . . think about the possibility of *extreme* political inequality, involving great political influence by a very small number of wealthy individuals. We argue that it is useful to think about the US political system in terms of oligarch.[2]

What are we to make of a (perhaps startling) claim such as this? How do we know it's true? Should we accept it?

[1] Jeffrey A. Winters and Benjamin I. Page, "Oligarchy in the United States?," *Perspective on Politics*, 7 (December 2009): 731–51.

[2] Ibid., 744. (Emphasis in original.) Also see Jeffrey A. Winters, *Oligarch* (New York: Cambridge University Press, 2014).

As the title of our book and the first chapter suggest, we have confidence in a statement like Winters and Page's *if* they arrive at their (tentative) conclusion through **empiricism**. This term is perhaps best explained by reference to an old joke.

Three baseball umpires were discussing their philosophy of calling balls and strikes. The first umpire says, "I call 'em as I see 'em." The next one replies, "That's nothing. I call 'em as they *are*." Finally, the third chimes in," Oh yeah! Well, they ain't *nothing* until I call them."

We put aside Umpires 1 and 3 until later in the chapter. For now let's concentrate on the second one. We call him a strict or strong empiricist. He believes there are in fact things like balls and strikes, and he can always tell the difference by merely looking at the pitches as they are thrown. He believes no interpretation is necessary; the facts (the pitches) speak for themselves, the umpire simply reports on where the ball travels, nothing more, nothing less. The teams, players, managers, fans have no bearing, he believes, on his judgments.[3]

An empiricist, in other words, uses observation to judge the tenability of arguments. A political science "umpire" demands that data and measurements support whatever point is being made. Statements can be believed and accepted to the extent that they are derived from empirical or observational evidence. If, on the other hand, their "truthfulness" depends on belief, authority, or faith instead of "hard data," they are set aside for philosophers and others to evaluate.

Empiricism is an ideal. Most who adopt this methodology would admit that personal judgment plays a part in their research—they are perhaps closer to the first umpire, who calls the game as he "sees it." But so important is empiricism that we

[3] During his Senate confirmation hearing, Chief Justice John Roberts came close to capturing the essence of the empirical viewpoint when he told the committee, "Judges and justices are servants of the law, not the other way around. Judges are like umpires. Umpires don't make the rules; they apply them." He added, "My job is to call balls and strikes and not to pitch or bat." CNN.com, September 12, 2005. Accessed June 3, 2015. Available at http://www.cnn.com/2005/POLITICS/09/12/roberts.statement. In other words, judges "see" the law and the facts of a case as they are. Judiciary Committee chair Joe Biden, however, challenged Justice Roberts on his umpire analogy: "So, as much as I respect your metaphor, it's not very apt, because you get to determine the strike zone. . . . Your strike zone . . . may be very different than another judge's view." *Washington Post*, "Transcript: Day Two of the Roberts Confirmation Hearings," September 13, 2005. Accessed January 10, 2015. Available at http://www.washingtonpost.com/wp-dyn/content/article/2005/09/13/AR2005091300979.html. In other words, the senator believes judges may act like Umpire 3, who in a sense "constructs" reality in his own way.

need to take a detour to clarify why many political scientists prefer this methodology over other ways of obtaining knowledge. Although not everyone agrees, it does seem to have a "privileged" place in the discipline, and we need to explore its philosophical basis. This leads us to a discussion of the scientific method.[4]

Although empiricism does have a dominant place in contemporary political science, we stress that it has its share of critics, and we certainly don't maintain that it is the only or even the best way to study politics. There is plenty of room, we believe, for different research stances. Nor do we believe that quantitative analysis is superior to qualitative studies. (In practice, most research contains a mixture of both.) Proponents of alternatives work under many different labels, so we simply classify them as *nonempiricists.*[5]

Elements of Empiricism

What, then, distinguishes the empirical or scientific approach? In our daily lives we "know" things in many different ways. We know, for example, that water boils at 212 degrees Fahrenheit and that a virus causes Ebola. We also may "know" that liberals are "weaker" on national defense than are conservatives, or that democracy is "better" than dictatorship. In some cases, we know something because we believe what we read in the newspaper or hear on the radio. In other cases, we believe it because of personal experience or because it appears to be consistent with common sense or is what a trusted authority told us.

Modern political science, though, relies heavily on one kind of knowledge: knowledge obtained through objective observation, experimentation, and logical reasoning.[6] This way of knowing differs greatly from information derived from myth, intuition, faith, common sense, sacred texts, and the like. It has certain characteristics that these other types of knowledge do not completely share. The ultimate goal of scientific research, which is not always attained, is to use its results to construct theories that explain political phenomena.[7]

[4] It might be more accurate to use the words "scientific methods," since to define what is and what is not science is a notoriously tricky task, and not everyone agrees on an exact definition.

[5] Those who follow the philosophy of social science, or epistemology, know that naming the sides in these methodological debates is virtually impossible. Someone we might label a *nonempiricist* might very well foreswear the tag. We are just attempting to sort out tendencies.

[6] Careful readers will note that we are combining all sorts of activities under one label. Specialists in one method or another often call themselves different things to emphasize the kind of research they do. For instance, those who rely on deductive reasoning and do not spend much time observing the world often refer to themselves as "formal modelers" or "rational choice theorists."

[7] Whether or not political science or any social science can find causal laws is very much a contentious issue in philosophy. See, for instance, Alexander Rosenberg, *The Philosophy of Social Science,* 3rd ed. (Boulder, Colo.: Westview, 2007).

Scientific knowledge exhibits several characteristics. Most important, scientific knowledge depends on **verification**. That is, our acceptance or rejection of a statement regarding something "known" must be influenced by observation.[8] Thus, if we say that people in the upper classes have more political power than members of the lower strata, we must be able to provide tangible evidence to support this statement. People often state opinions (beliefs) as if they were a matter of fact in rhetorical sentences, as in "No tax hike ever created a job." Without verification, this is not an empirical statement.

The contention cannot be accepted simply because someone said so or our instinct tells us so. The empirical nature of scientific knowledge distinguishes it from mystical knowledge. In the latter case, only "true believers" are able to observe the phenomena that support their beliefs, and observations that would disprove their beliefs are impossible to specify. Knowledge derived from superstition and prejudice is usually not subjected to empirical verification, either. Superstitious or prejudiced persons are likely to note only phenomena that reinforce their beliefs, while ignoring or dismissing those that do not. Thus, their knowledge is based on selective and biased experience and observation.

On the flip side, some philosophers of science insist that a key characteristic of scientific claims is **falsifiability**, meaning the statements or hypotheses can in principle be rejected in the face of contravening empirical evidence. A claim not refutable by any conceivable observation or experiment is nonscientific. (How does one empirically refute "God is great"?) In this sense, the findings of science are usually considered tentative, because they are "champions" only so long as competing ideas do not upend them. Indeed, the philosopher Karl Popper argued that scientists should think solely in terms of invalidating or falsifying theories, not proving them.[9]

In·view of the importance of verification and falsification, researchers must always remain open to alterations and improvements of their research. To say that scientific knowledge is provisional does not mean that the evidence accumulated to date can be ignored or is worthless. It does mean, however, that future research could significantly alter what we currently believe. In a word, scientific knowledge is *tentative* and because of this property, empirical research is thought to be self-corrective.

Sometimes efforts to investigate commonsense knowledge have surprising results. For example, given the United States' high levels of literacy, the emergence of mass communications, the development of modern transportation networks, and the steady expansion of voting rights for the last two hundred years, we might assume

[8] Ibid., 107.

[9] The most ardent proponent of the idea that science really amounts to an effort to falsify (not prove) hypotheses and theories is Karl Popper. See, for example, *The Logic of Scientific Discovery* (New York: Basic Books, 1959).

that participation in national elections would be high and that it would increase as time goes by. But, as it turns out, neither of these conditions holds. Lots of evidence indicates that half or more of eligible Americans regularly skip voting, and that the number doing so may be increasing despite all the economic and civic progress that has been made.

Scientific knowledge is supposedly "value-free." Empiricism addresses what is, what might be in the future, and why. It does not typically address whether or not the existence of something is good or bad, although it may be useful in making these types of determinations. Political scientists use the words *normative* and *nonnormative* to express the distinction. Knowledge that is evaluative, value-laden, and concerned with prescribing what ought to be is known as **normative knowledge**. Knowledge that is concerned not with evaluation or prescription but with factual or objective determinations is known as **nonnormative knowledge**. Most scientists would agree that science is (or should attempt to be) a nonnormative enterprise.

This is not to say that empirical research operates in a valueless vacuum. A researcher's values and interests, which are indeed subjective, affect the selection of research topics, periods, populations, and the like. A criminologist, for example, may feel that crime is a serious problem and that long prison sentences deter would-be criminals. He or she may therefore advocate stiff mandatory sentences as a way to reduce crime. But the researcher should test that proposition in such a way that personal values and predilections do not bias the results of the study. And it is the responsibility of other social scientists to evaluate whether or not the research meets the criteria of empirical verification. Scientific principles and methods of observation thus help both researchers and those who must evaluate and use their findings. Note, however, that within the discipline of political science, as well as in other disciplines, the relationship between values and scientific research is frequently debated. We have more to say about this subject later in the chapter.

An additional characteristic of scientific knowledge helps to identify and weed out prejudices (inadvertent or otherwise) that may creep into research activities.[10] Scientific knowledge must be **transmissible**—that is, the methods used in making scientific discoveries must be made explicit so that others can analyze and replicate findings. The transmissibility of scientific knowledge suggests "science is a social activity in that it takes several scientists, analyzing and criticizing each other, to produce more reliable knowledge."[11] To accept results, people must know what data were collected and how they were analyzed. A clear description of research procedures allows this independent evaluation. It also permits other scientists to collect the same information and test the original propositions themselves. If researchers

[10] Alan C. Isaak, *Scope and Methods of Political Science*, 4th ed. (Homewood, Ill.: Dorsey, 1985), 30.

[11] Ibid., 31.

HELPFUL HINTS

Types of Assertions

It is sometimes tricky to tell an empirical statement from a normative one. The key is to infer the author's intention: Is he or she asserting that something is simply the way it is, no matter what anyone's preference may be? Or is the person stating a preference or desire? Sometimes normative arguments contain auxiliary verbs, such as *should* or *ought,* which express an obligation or a desire. Empirical arguments, by contrast, often use variations of *to be* or direct verbs to convey the idea that "this is the way it really is in the world." Naturally, people occasionally believe that their values are matters of fact, but scientists must be careful to keep the types of claims separate. Finally, people often state opinions (beliefs) as if they were a matter of fact in rhetorical sentences, as in "No tax hike ever created a job."

When reading research reports or (even more important) when following political discussions in the media, on the Internet, or on the campaign trail, try to keep in mind that statements that seem to be of the same type can be surprisingly different:

- Empirical: A verifiable assertion of "what is"
- Normative: An assertion of "what should be"
- Rhetorical: A statement to the effect that "my belief is a fact"

The Web site and workbook contain examples.

Check out more Helpful Hints at **edge.sagepub.com/johnson8e**

use the same procedures but do not replicate the original results, something is amiss and the reasons for the discrepancy must be found. Until then, both sets of results are suspect.

This idea leads to another characteristic of empirical knowledge: it is **cumulative**, in that both substantive findings and research techniques are built upon those of prior studies. As Isaac Newton famously observed of his own accomplishments, "I have stood on the shoulders of giants." He meant that the attainment of his revolutionary insights depended in part on the knowledge other scientists had generated in the previous decades and centuries. The process of constantly testing and refining prior research produces an accumulated body of knowledge. (You'll see examples of this fact in chapter 3, which explains literature reviews.)

Another important characteristic of scientific knowledge is that it is general, or applicable to many rather than just a few cases. Advocates of the scientific method argue that knowledge that describes, explains, and predicts many phenomena or a set of similar occurrences is more valuable than knowledge that addresses a single phenomenon.[12] For example, the knowledge that states with easier voter registration systems have higher election turnout rates than do states with more difficult systems is preferable to the knowledge that Wisconsin has a higher turnout rate than does Alabama. Knowing that party affiliation strongly influences many voters' choices among candidates is more useful knowledge to someone seeking to understand elections than is the simple fact that John Doe, a Democrat, voted for a Democratic candidate for Congress in 2006. The knowledge that a state that has a safety inspection program has a lower automobile fatality rate than another state, which does not, is less useful information to a legislator considering the worth of mandatory inspection programs than is the knowledge that states that require automobile inspections experience lower average fatality rates than those that do not.

The empirical approach thus strives for empirical generalizations, statements that describe relationships between particular sets of facts.[13] For example, the assertion that positive campaigns lead to higher voter turnout than do those that are characterized by mudslinging and name-calling is intended to summarize a relationship that holds in different places and at different times.

Another characteristic of scientific knowledge is that it is **explanatory**; that is, it provides a systematic, empirically verified understanding of why a phenomenon occurs. In scientific discourse, the term *explanation* has various meanings, but when we say that knowledge is explanatory, we are saying that a conclusion can be derived (logically) from a set of general propositions and specific initial conditions. The general propositions assert that when things of type X occur, they will be followed by things of type Y. An initial condition might specify that X has in fact occurred. The observation of Y is then explained by the conjunction of the condition and the proposition. The goal of explanation is, sometimes, to account for a particular event—the emergence of terrorism, for example—but more often it is to explain general classes of phenomena such as wars or revolutions or voting behavior.

Explanation, then, answers "why" and "how" questions. The questions may be specific (e.g., "Why did a particular event take place at a particular time?") or more general (e.g., "Why do upper-class people vote more regularly than, say, blue-collar

[12] It may be tempting to think that historians are interested in describing and explaining only unique, onetime events, such as the outbreak of a particular war. This is not the case, however. Many historians search for generalizations that account for several specific events. Some even claim to have discovered the "laws of history."

[13] Isaak, *Scope and Methods,* 103.

workers?"). Observing and describing facts are, of course, important. But most political scientists want more than mere facts. They are usually interested in identifying the factors that account for or explain human behavior. Studies of turnout are valuable because they do more than simply describe particular election results; they offer an explanation of political behavior in general.

An especially important kind of explanation for science is that which asserts *causality* between two events or trends. A causal relationship means that in some sense, the emergence or presence of one condition or event will always (or with high probability) bring about another. Causation implies more than that one thing is connected to or associated with another. Instead, it means one *necessarily* follows the other. Chapter 1 touched on the issue of why economic inequality appears to be increasing in the United States. Some political scientists, for instance, believe that "de-unionzation" (the weakening of organized labor) has led to (caused) an increase in inequality in the United States. But is there, in fact, a causal connection, or is the relationship merely fortuitous? Statements asserting cause and effect are generally considered more informative and perhaps more useful than ones simply stating that an unexplained connection exists. But they are difficult to establish.

In this vein, explanatory knowledge is also important because, by offering systematic, reasoned anticipation of future events, it can be predictive. Note that prediction based on explanation is not the same as forecasting or soothsaying or astrology, which do not rest on empirically verified explanations. An explanation gives scientific reasons or justifications for why a certain outcome is to be expected. In fact, many scientists consider the ultimate test of an explanation to be its usefulness in prediction. Prediction is an extremely valuable type of knowledge, since it may be used to avoid undesirable and costly events and to achieve desired outcomes. Of course, whether or not a prediction is "beneficial" is a normative question. Consider, for example, a government that uses scientific research to predict the outbreak of popular unrest but uses the knowledge not to alleviate the underlying conditions but to suppress the discontented with force.

In political science, explanations rarely account for all the variation observed in attributes or behavior. So exactly how accurate, then, do scientific explanations have to be? Do they have to account for or predict phenomena 100 percent of the time? Most political scientists, like scientists in other disciplines, accept probabilistic explanation, in which it is not necessary to explain or predict a phenomenon with 100 percent accuracy.

Scientists also recognize another characteristic of scientific knowledge: **parsimony**, or simplicity. Suppose, for instance, two researchers have developed explanations for why some people trust and follow authoritarian leaders. The first account mentions only the immediate personal, social, and economic situation of the individuals, whereas the second account accepts these factors but also adds deep-seated psychological states stemming from traumatic childhood experiences. And imagine that both provide equally compelling accounts and predictions of behavior. Yet, since the first relies on fewer explanatory factors than does the second, it will generally be the preferred explanation, all other things being equal. This is the principle of Ockham's razor, which might be summed up as "keep explanations as simple as possible."

The Importance of Theory

The accumulation of related explanations sometimes leads to the creation of a **theory**—that is, a body of statements that systematize knowledge of and explain phenomena. Two crucial aspects of empirical theory are (1) that it leads to specific, testable predictions, and (2) that the more observations there are to support these predictions, the more the theory is confirmed.

An Example: Proximity Theory of Voting

To clarify some of these matters, let us take a quick look at an example. The "proximity theory of electoral choice" provides a concise explanation for why voters choose parties and candidates.[14] Superficially, the theory may seem simplistic. Its simplicity can be deceiving, however, for it rests on many years of multidisciplinary research[15] and involves considerable sophisticated thinking.[16] But essentially the theory boils down to the assertion that people support parties and candidates who are "closest" to them on policy issues. Furthermore, this theory would predict that candidates will try to position themselves so that they are closer to more voters than are their opponents.

Take a particularly simple case. Suppose we consider the immigration debate. Positions on this issue might be arrayed along a single continuum running from, say, "All undocumented immigrants should have a path to citizenship" to "All undocumented immigrants should be deported" (see figure 2-1). Proximity theorists believe that both voters and candidates (or parties) can be placed or located on this scale and, consequently, that the distances or proximities between them (voters and

[14] Many varieties of this theory exist, but they share the components presented here.

[15] Anthony Downs, an economist, provided one of the first explications of the theory in *An Economic Theory of Democracy* (New York: Harper & Row, 1957). His ideas in turn flowed from earlier economic analyses. See, for example, Harold Hotelling, "Stability in Competition," *Economic Journal* 39, no. 153 (1929): 41–57. Available at http://www.edegan.com/pdfs/Hotelling (1929)—Stability in competition.pdf

[16] See James Enelow and Melvin Hinch, *The Spatial Theory of Voting: An Introduction* (New York: Cambridge University Press, 1984).

HELPFUL HINTS

Keep Terms Straight: *Theory* versus *Fact*

It's common to hear people say, "Evolution is a theory, not a fact," as if being a "theory" is grounds for doubt. Surprisingly perhaps, most scientists would reply, "So what if it is a theory?" And they would immediately add that evolution is a particular kind of theory, a *scientific* theory.* A scientific theory has, among other attributes, the property that its claims can be falsified by empirical observation and testing. The ideas and predictions of evolutionary theory have been repeatedly tested and confirmed by scientific methods and standards, and evolutionary theory today (it's sometimes called the "modern synthesis") rests on a solid body of confirmed evidence. So nearly every scientist accepts it as the more or less valid account of, say, human ancestry. But being verified is not the same as being "proved." So no, evolution is not a proven fact but a set of well-supported findings. Scientists will gladly abandon Darwinism if and when a better scientific explanation comes along.

If one relies on a different epistemology (or method for determining what is true), such as faith or the advice of trusted people, then the fact that evolutionary theory has the support of science may be beside the point. One might still doubt that humans and chimpanzees share a common ancestor. In so doing one is using a different—not necessarily inferior—methodology for deciding what is true. But this method is *not* based on scientific principles.

*The biologist Richard Dawkins writes in *The Greatest Show on Earth* (New York: Free Press, 2009): "Evolution is a fact" (p. 8). Despite appearances to the contrary, Dawkins knows evolution is a theory and clearly illustrates how scientists use the word *theory* in a scientific context, where its meaning may differ from that of common usage.

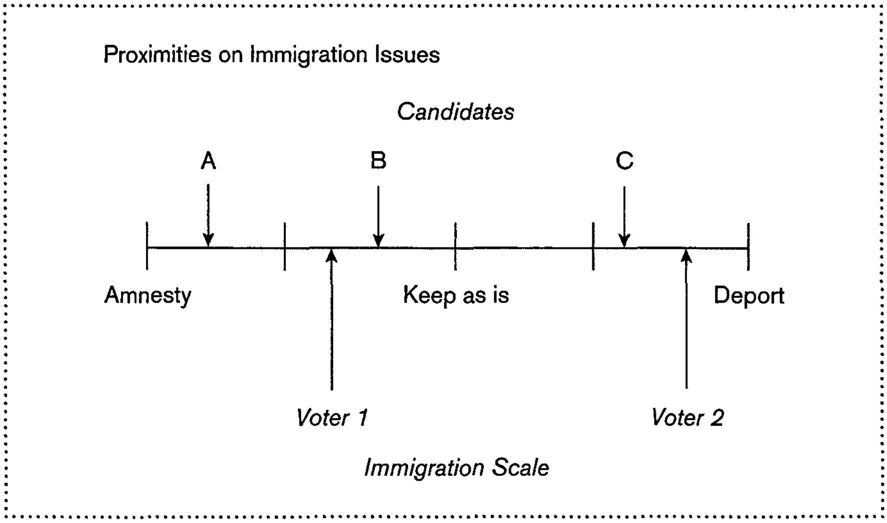
Check out more Helpful Hints at **edge.sagepub.com/johnson8e**

candidates) can be compared. The theory's prediction is straightforward: an individual votes for the candidate to whom he or she lies closest on the continuum.[17]

To expand a bit, theorists in this camp argue (1) that analysts using proper measurement techniques can position both issues and candidates on scales that show how "close" they are to each other and to other objects, and (2) that voters vote for

[17] This expectation assumes that immigration is important to the voter—that there is not some other issue that is more important that may cause the voter to prefer another candidate.

FIGURE 2-1 **Proximities on Immigration Issues**



candidates who are closest (most proximate) to themselves on such scales. People choose nearby candidates out of their desire to maximize utility, or the value that results from one choice over another. Knowing this fact, candidates adjust their behavior to maximize the votes they receive. Adjusting behavior means not only taking or moving to positions as close as possible to those of the average or typical voter (the so-called median voter) but also, if and when necessary, obscuring one's true position (that is, following a strategy of ambiguity).[18] Figure 2-1, for instance, shows that Voter 1's position is closest to Candidate B's; therefore, Voter 1 would presumably vote for that person. Similarly, Voter 2 would prefer Candidate C. Note also that Candidate A could attract Voter 1's support by moving closer to the middle, perhaps by campaigning on an "amnesty-only-for-children-of-illegal-immigrants" platform.

The proximity theory has many of the characteristics of an empirical theory. Note that it does not take a stance for or against one side or the other in the immigration debate. Rather, it explains why things happen as they do, and it offers specific and testable predictions. It is also an implicitly causal theory in that it hypothesizes that the desire to maximize utility "causes" voters to vote for specific candidates. It is general since it claims to apply to any election in any place at any time. As such, it

[18] Kenneth Shepsle, "The Strategy of Ambiguity: Uncertainty and Electoral Competition," *American Political Science Review* 66, no. 2 (1972): 555–68.

provides a much more sweeping explanation of voting than a theory that uses time- and place-bounded terms such as "the 2014 gubernatorial election in Pennsylvania." In addition, it provides a parsimonious or relatively simple account of candidate choice. It does not invoke additional explanatory factors such as psychological or mental states, social class membership, or current economic conditions to describe the voting act. Most important, although the proximity theory rests on considerable formal (and abstract) economic and decision-making reasoning, it puts itself on the line by making specific empirical predictions, which can be checked by asking voters (1) their positions on immigration and (2) how they voted.

As a theory, it incorporates or uses numerous primitive or undefined terms such as *issue, candidate,* and *utility.* These words and concepts may have well-accepted dictionary meanings, but the theory itself takes their common understanding for granted. When a theory is challenged, part of the dispute might involve slightly divergent interpretations of these terms. At the same time, the theory makes explicit various other assumptions. It assumes, among other things, that a researcher can place individuals on issue dimensions, that people occupy these positions for reasonably long time periods, that voters are rational in that they maximize utility, and that candidates have objective positions on these issues.[19] Moreover, by assumption, certain possibilities are not considered. The theory does not delve into the question of whether or not a person holds a "correct" position on the scale, given his or her objective interests. Finally, to test the proximity or spatial idea, researchers assume that one can assign individuals meaningful spatial positions by asking certain kinds of questions on surveys or polls.[20] This may be a perfectly reasonable assumption (we touch on that matter in chapter 10), but it is an assumption nevertheless. Still, spatial modelers, as those who use proximity theory are called, go to great lengths to define and explain key concepts. How *distance* is defined is a serious matter because different definitions can lead to different substantive conclusions.[21]

[19] As an example, see Anders Westholm, "Distance versus Direction: The Illusory Defeat of the Proximity Theory of Electoral Choice," *American Political Science Review* 91, no. 4 (1997): 870.

[20] Here is an example: "Please look at . . . the booklet. Some people believe that we should spend much less money for defense. Suppose these people are at one end of a scale, at point 1. Others feel that defense spending should be greatly increased. Suppose these people are at the other end, at point 7. And, of course, some other people have opinions somewhere in between, at points 2, 3, 4, 5 or 6." See *American National Election Study (ANES) 2004 Codebook* (2006). Available at the Survey Documentation and Analysis, University of California–Berkeley, Web site: http://sda.berkeley.edu/D3/NES2004public/Doc/nesO.htm

[21] The conceptualization of distance and other matters related to the proximity theory are debated in Westholm, "Distance versus Direction," 865–73; and Stuart Elaine Macdonald, George Rabinowitz, and Ola Listhaug, "On Attempting to Rehabilitate the Proximity Model: Sometimes the Patient Just Can't Be Helped," *Journal of Politics* 60, no. 3 (1998): 653–90.

The Explanatory Range of Theories

Theories are sometimes described by their explanatory range, or the breadth of the phenomena they purport to explain. Usually one does not have a theory of "why Barack Obama won reelection in 2012." (It is, of course possible to find several theories that account for this particular outcome. But note that the 2012 election results are an instance, or "token," of the kind of event with which these theories deal.) Instead, a good theory of electoral outcomes presumably pertains not only to a specific presidential contest but also to other elections in other times and places.

In the social sciences, so-called narrow-gauge or middle-range theories pertain to limited classes of events or behaviors, such as a theory of voting behavior or a theory about the role of revolution in political development.[22] Thus, a theory of voting may explain voter turnout by proposing factors that affect people's perceptions of the costs and benefits of voting: socioeconomic class, degree of partisanship, the ease of registration and voting laws, choices among candidates, availability of election news in the media, and so forth.[23] Global or broad-range theories, by contrast, claim to describe and account for an entire body of human behavior. A really general theory, for example, might attempt to account for increases or decreases in economic inequality in any society at any time.[24] In short, theories play a prominent role in natural and social sciences because they provide general accounts of phenomena.[25] Other things being equal, the broader the range of the things to be explained, the more valuable the theory.

A Brief Overview of the Empirical Research Process

So what exactly is the empirical or scientific research process? In reality, no scientist in the field or laboratory adheres to a prescribed set of steps like someone following a script. Scientists rely not just on formal procedures but also on intuition, imagination, and even luck at times. Nevertheless, we may conceptualize what they do by identifying the underlying logic of their activities. Here is an idealization of a scientific research program.

[22] A good example is Theda Skocpol, *States and Social Revolutions: A Comparative Analysis of France, Russia, and China* (New York: Cambridge University Press, 1979).

[23] See Raymond E. Wolfinger and Steven J. Rosenstone, *Who Votes?* (New Haven, Conn.: Yale University Press, 1980).

[24] A good example is Thomas Picketty, *Capitalism in the Twenty-first Century* (Boston: Belknap/Harvard, 2014).

[25] Isaak, *Scope and Methods*, 167.

Development of an Idea to Investigate or a Problem to Solve

A scientist gets topics from any number of sources, including literature about a subject, a general observation, an intuition (or hunch), the existence of conflicts or anomalies in reported research findings, and the implications of an established theory. For example, a report on income inequality may indicate that it varies considerably from country to country or that it is increasing. A logical response would be to ask "why?" As another instance, consider newspaper accounts that suggest that evangelical Christians tend to support conservative candidates because of "moral values." Several research questions are raised by these accounts: Do evangelicals base their choices of candidates on their proximity to candidates' positions on moral issues while other voters base their choice on other types of issues? Do evangelicals turn out to vote more in elections where there are distinct differences between candidates on moral issues than in elections where the differences are small?

Hypothesis Formation

After selecting a topic, an investigator tries to translate the idea or problem into a series of specific hypotheses. As we see in chapter 4, hypotheses are tentative statements that, if confirmed, show how and why one thing is related to another or why a condition comes into existence. These statements have to be worded unambiguously and in a way that their specific claims can be evaluated by commonly accepted procedures. After all, one of the requirements of science is for others to be able to independently corroborate a discovery. If assertions are not completely transparent, how can someone else verify them? In the preceding example, we might hypothesize that evangelical Christians are more likely than others to base their vote on candidates' positions on moral issues.

"Data" Collection

This is where the rubber meets the road: the essence of science comes in the empirical testing of hypotheses through the collection and analysis of data. Consider the case of religion and voting just mentioned. We need to define clearly the concepts of *moral values* and *evangelical Christian*. We might, for instance, tentatively identify evangelicals as people who adhere to certain Christian denominations and moral values as attitudes toward abortion and gay marriage. It would be possible (but not necessarily easy) to write a series of questions to be administered in a survey or a poll to elicit this information. Only when concepts are defined and decisions made about how to measure them can data collection and analysis begin.

Interpretation and Decision

At some point the investigator has to determine whether or not the observed results are consistent with the hypotheses. Though simple in principle, judging how well data support scientific hypotheses is usually not an easy matter. Suppose, for example, we find that 75 percent of evangelical Christians opposed gay marriage and 90 percent of these individuals voted for a House candidate in 2014 who opposed gay marriage. So far, so good. But suppose, in addition, that 70 percent of non-evangelicals also opposed gay marriage and that more than 90 percent of these people also voted for House candidates opposed to gay marriage in the same election. It appears that attitudes might be affecting voting, but the data do not necessarily establish a connection between religious preference and whether or not votes are based on moral issues. Weighing quantitative or statistical evidence requires expertise, practice, and knowledge of the subject matter, plus good judgment (and this skill is often difficult to teach). Still, chapters in this book are devoted to showing ways to make valid inferences about tenability of empirical hypotheses.

Modification and Extension

Depending on the outcome of the test, one can tentatively accept, abandon, or modify the hypothesis. If the results are favorable, it might be possible to derive new predictions to investigate. If, however, the data do not or only very weakly support the hypothesis, it will be necessary to modify or discard it. Let us stress here that negative results—that is, those that do not support a particular hypothesis—can still be both interesting and beneficial.[26] As we suggested earlier, some scholars, such as Popper, believe that science advances by disproving claims, not by accepting them. Consequently, a valuable contribution to science can come from disconfirming widely held beliefs, and the only way to do that is to replicate or reinvestigate the research upon which the beliefs rest. The key is not so much the result of a hypothesis test but how substantively important the hypothesis is to begin with.

Reactions to the Empirical Approach: Practical Objections

Empirical research problems arise because many important concepts are abstract or have many meanings or are value-laden. Chapter 1 showed that an idea as seemingly straightforward as "the number of eligible voters" can present problems

[26] An often-remarked-on characteristic of scholarly journals is that they tend to report mostly positive findings. An article that shows "X is related to Y" may be more likely to be accepted for publication than one that asserts "X is *not* related to Y." Whether or not a "negative result" makes a significant contribution to knowledge depends on the importance of the original claim. Suppose that a team of psychologists found that "love and marriage" really do *not* "go together." That would be worth publishing.

that affect our substantive conclusions about how civic minded Americans are. Or finding an adequate definition of "economic inequality" can be difficult. Should we be looking at individuals or households? Should we use annual income—calculated before taxes, after taxes, or after adding to individual or household income publicly supplied in-kind benefits such as health care or job retraining? Or should we try to measure net wealth (assets minus debts)? The following chapters take up some of these questions.

Furthermore, political scientists must face the fact that human behavior is complex, perhaps even more complex than the subject matter of other sciences (genes, subatomic particles, insects, and so on). Complexity has been a significant obstacle to the discovery of general theories that accurately explain and predict almost every kind of behavior. After all, developing a theory with broad applicability requires the identification and specification of innumerable variables and the linkages among them. Consequently, when a broad theory is proposed, it can be attacked on the grounds that it is too simple or that too many exceptions to it exist. Certainly to date no empirically verified generalizations in political science match the simplicity and explanatory power of Einstein's famous equation $E = mc^2$.[27]

There are still other obstacles. The data needed to test explanations and theories may be extremely hard to obtain. Indeed, often the potentially most informative data are totally unavailable. People with the needed information, for example, may not want to release it for political or personal reasons. Pollsters, for instance, find refusal to answer certain questions, such as those designed to measure attitudes toward ethnic groups, to be a major problem in gauging public opinion. Similarly, some experiments require manipulation of people. But since humans are the subjects, the researchers must contend with ethical considerations that might preclude them from obtaining all the information they want. Asking certain questions can interfere with privacy rights, and exposing subjects to certain stimuli might put the participants at physical or emotional risk. Tempting someone to commit a crime, to take an obvious case, might tell a social scientist a lot about adherence to the law but would be unacceptable nevertheless.

Self-Reflection and Individuality

Like any other organisms, humans are aware of their surroundings. They have the additional ability to empathize with others and frequently attempt to read others' minds. As John Medearis put it, "human beings—individually, but especially jointly—are self-interpreting and reflective, capable of assigning meanings to their actions and revising these meanings recursively."[28] Observations of this sort led many social scientists and philosophers to question whether or not the scientific

[27] For further discussion of complete and partial explanations, see Isaak, *Scope and Methods,* 143.

[28] Medearis, review of *Perestroika!,* 577.

method can be applied to the study of something as intrinsically language based as politics. This doubt appears later in the chapter, when we discuss interpretation versus explanation. In the meantime, let us point to a practical problem. Since humans are self-reflective and empathetic creatures, they often anticipate a researcher's goals and adjust their actions accordingly (e.g., "The investigator seems to favor immigration reform, so I will too").

When it comes to studying political behavior such as voting or decision making, another difficulty arises. Many experiments in science assume that the entities under investigation are for all intents and purposes identical and, hence, can be interchanged without fear of compromising the conclusion. An iron ion (Fe^+) from one source is as good as another from somewhere else (no matter where in the universe) when it comes to studying iron's reaction with oxygen. But can the same be said of humans? Consider a political scientist who wants to investigate the effects of negative campaign advertising on attitudes. Suppose Jane and Mary are subjects in a study. We cannot assume that they will react to the experimental stimuli exactly the same way, even though they are the same age, gender, political persuasion, and so forth.

Social scientists have to get around this problem by using groups or samples of individuals and then examining the *average* effect of the stimulus. Any generalization that results has the form: given subjects with characteristics A, B, . . . , X (the stimulus) *on average* affects Y (the response) by *approximately* N units. In other words, sometimes the basic units under the scientist's microscope can be considered pure, even if they are complex molecules, but not so in political science. The objects political scientists study are multifaceted and conscious beings with a volition of their own who often change opinions and behaviors; thus, statements about them must necessarily be tentative, general, and time bound.

Finally, there is the inescapable subjectivity of politics. We provide an example that bedevils research into the studies of power. Most political scientists would agree that, if an oligarchy exists in the United States, it should at a minimum make or heavily influence key policy decisions. The problem is, how does one objectively identify "key" policies? Should the choice be left to the judgment of the researcher or knowledgeable/informed experts? Or are there concrete indicators or measures of importance? Suppose we want to class decision A as "important." On what grounds do we make the assignment? The number of people A affects? Its cost? The number of times it is mentioned in the press? Its length in legal codes? The number of times it is litigated? Any or all of these might be useful. But for a variety reasons none of these may capture the significance (or lack of significance) of a decision. Importance often comes from how people interpret or understand policy A, and understanding of this sort, many assert, lies beyond the scope of empirical sciences.[29]

[29] For an effort to objectively measure policy importance, see David Mayhew, *Divided We Govern: Party Control, Lawmaking, and Investigations, 1946–2002,* 2nd ed. (New Haven, Conn.: Yale University Press, 2005).

All of these claims about the difficulty of studying political behavior scientifically may have merit. Yet they can be overstated. Consider, for example, that scientists studying natural phenomena encounter many of the same problems. Physicists cannot directly observe elementary particles such as quarks. Nor can astronomers and geologists carry out experiments on most of the phenomena of greatest interest to them. Indeed, they cannot even visit many of the places they study most intensively, like other planets or the center of the Earth. And what can be more complex than biological organisms and their components, which consist of thousands of compounds and chemical interactions? Stated quite simply, it is in no way clear that severe practical problems distinguish political science from any of the other sciences.

Is Political Science Trivial or Irrelevant?

The empirical approach in political science, with its advent in earnest in the 1960s, seemed to bring with it all the accoutrements of rigorous natural sciences: equations and mathematical models, statistical analysis, instrumentation and quantification, computers and electronic databases, esoteric concepts (e.g., "multidimensional issue spaces"). Yet practically from the moment the empirical or scientific perspective arrived on the scene, doubters and skeptics appeared. Among other complaints, they pointed to the trivial nature of some of the "scientific" findings and applications. Common sense would have told us the same thing, they argued. Moreover, and far worse in some people's minds, the empirical approach with its emphasis on quantification seemed to become more and more irrelevant to a practical understanding of government, a concern that persists to this day:

> Academics have followed the architectonic path of turning the study of politics into a theoretical pursuit unconcerned with the needs of and far removed from the understanding of the ordinary citizen or political leader. No one reading the last dozen issues of the *American Political Science Review* would find much that would provide an answer to the most fundamental of all political questions: "What is to be done?"[30]

Of course, as we explained earlier, there is a difference between intuition and scientific knowledge. To build a solid base for further research and accumulation of scientific knowledge in politics, commonsense knowledge must be verified empirically and, as is frequently the case, discarded when wrong. Still, "scientism" left many political scientists dismayed.

[30] Stephen B. Smith, "Political Science and Political Philosophy: An Uneasy Relation," *PS: Political Science and Politics* 33, no. 2 (2000): 189. Or, "We proceed with a two-fold working hypothesis: (1) Academic political science has very little awareness of the knowledge about politics held by practitioners, and (2) Political science is increasingly limited to a highly abstract understanding of politics. . . . We subscribe . . . to the view that academics have limited understanding about the practical work of politics and governance. The academic understanding expressed in concepts, models, and theories is abstract and usually innocent of the nuances regularly experienced by practitioners"; see John R..Petrocik and Frederick T. Steeper, "The Politics Missed by Political Science," *The Forum* 8, no. 3 (2010): 1.

A more serious criticism of the scientific study of politics is that it leads to a failure to focus enough scholarly research attention on important social issues and problems. Some critics contend that, in the effort to be scientific and precise, political science overlooks the moral and policy issues that make the discipline relevant to the real world. Studies rarely address the implications of research findings for important public policy choices or political reform. In other words, the quest for a scientific knowledge of politics has led to a focus on topics that are quantifiable and relatively easy to verify empirically but that are not related to significant, practical, and relevant societal concerns.[31] In the late 1960s and later in 2000, well-publicized "revolts" against hard-core empiricism took place. After all, to say "I'm only concerned with facts" may be to turn a blind eye to human suffering and injustice.

These considerations take us back to our umpires. Can researchers really emulate Umpire 2 (the strict empiricist) who claims to call "'em as they *are*"? Many think not. Political scientists, having been exposed to decades of philosophizing about limitations and problems of the "scientific method," probably now admit to being like Umpire 1 and call balls and strikes as they *see* them. This doesn't mean their research is totally subjective or a matter of opinion; but it is, they realize, so contingent on time, place, language, and culture that finding scientific laws and truths of politics is problematic. Instead of calling them hard-nosed empiricists, we might better call today's political scientists modest or constrained empiricists.

Competing Points of View

As widely accepted and useful as science has become in modern times, serious philosophers and social scientists have challenged these premises. We cannot explain all of their objections here, but the essence of their argument is that certain aspects of human life are simply not amenable to systematic and objective analysis. More important, an uncritical faith in realism, objectivity, and material causality is unwarranted. We concentrate on two points:

1. Human actions cannot be explained scientifically but must be *interpreted* from the point of view of the actors. Meaning and understanding are the proper goals.
2. Social scientists have to realize that the world, far from having an independent existence that they observe directly, is partly *constructed* by observers themselves.

To oversimplify, we shall say these two viewpoints constitute "nonempiricism."

[31] See Charles A. McCoy and John Playford, eds., *Apolitical Politics: A Critique of Behavioralism* (New York: Thomas Y. Crowell, 1967).

HELPFUL HINTS

Assumptions of Empirical Research

- An empiricist (I-call-'em-as-they-are umpire) makes assumptions about methodology.
- *Realism*: There is a real world that exists independently of observers. (It's there even if we aren't there to see it.)
- *Materialism*: Only concrete and observable (if only indirectly) entities have causal efficacy.
- *Denial of supernatural causes*: Explanations of phenomena based on mysterious, unknowable, unobservable, "hidden" forces are unacceptable.
- *Regularity*: Natural phenomena (human behavior and institutions) exhibit regularities and patterns that can be revealed by reason and observation.
- *Verification and falsification*: Statements about the world must be verified or falsified by experience or data. (Don't take anything on faith alone.)
- *Irrelevance of preferences*: To the maximum extent, one's values and biases should not affect the decision to reject or accept an empirical claim.
- *Theory and causal explanation*: The goal of science is to create general, verified explanatory theories (even laws).

Check out more Helpful Hints at **edge.sagepub.com/johnson8e**

Interpretation

Some people question the empirical strategy because the subject matter, human institutions and activities, differs from the behavior of material objects such as atoms or stars, and these differences raise all sorts of complexities. One indicator of the inapplicability is that progress in developing and testing contingent causal laws has been agonizingly slow.[32] Moreover, both the methods and the content of the discipline have not come close to the exactitude and elegant sophistication of sciences such as biology or physics, and, consequently, nowhere can we find empirical generalizations with the level of precision and confirmation enjoyed by, say, the theories of relativity and evolution.

[32] See Alexander Rosenberg, *The Philosophy of Social Science*, 3rd ed. (Boulder, Colo.: Westview, 2007).

Skeptics argue that there are good reasons for this outcome. Since politics inescapably involves **actions**—that is, behavior that is done for reasons—and not mere physical movement, analyzing it brings up challenges not encountered in the natural sciences. Opponents of the empirical approach claim that scientific methods do not explain nearly as much about behavior as their practitioners think. The problem is that to understand human behavior, one must try to see the world the way individuals do. And doing so requires empathy, or the ability to identify and in some sense experience the subjective moods or feelings or thoughts of those being studied. Instead of acting as outside, objective observers, we need to "see" how individuals themselves view their actions. Only by reaching this level of understanding can we hope to answer "why" questions such as "Why did John still vote in the last election even though he was bombarded by countless attack ads on television, the Internet, radio . . . everywhere he turned?" The answers require the interpretation of behavior, not its scientific explanation in terms of general laws. In short, **interpretation** means decoding verbal and physical actions, which is a much different task than proposing and testing hypotheses.

Given this way of looking at the research task, some social scientists advocated stressing the interpretation or empathetic understanding of actions and institutions. One of the earliest and best-known proponents of this methodology was Clifford Geertz, an anthropologist, who felt that "man is an animal suspended in webs of significance he himself has spun. I take culture to be those webs, and the analysis of it to be therefore not an experimental science in search of law but an interpretive one in search of meaning."[33] As a simple example of the difference between empirical and interpretative approaches, take journalist James O'Toole's analysis of a close Pennsylvania US Senate election in 2010: "it's now a pretty close race, according to the polls and the body language of the campaigns."[34] Here he relies on both an empirical tool (polling) and intuition (the "body language of the campaigns").Those who closely follow electoral politics would perhaps agree that a minimum of interpretation and subjective analysis is always helpful.

Another way of looking at interpretation is to consider the concept of **social facts**. What exactly are things like political parties, elections, laws, and administrative regulations? In what sense are they real? They do not have same kind of material existence as atoms, bacteria, and mountains, but have an entirely subjective existence *only* in the minds of people living in a particular culture. One philosopher remarks that "minds create institutions. There would be no money or marriage or

[33] Clifford Geertz, *The Interpretation of Cultures* (New York: Basic Books, 1973), 5; see also following discussion, pp. 6–7.

[34] "Federal Spending Front and Center in Pa., Wash. Senate Races," *PBS NewsHour*, October 26, 2010. Available at http://www.pbs.org/newshour/bb/politics/july-dec10/campaign_10-26.html

private property without human minds to create these institutions."[35] How, then, should they be studied? The sociologist Émile Durkheim told his students to take them seriously: "the first and most basic rule [of social inquiry] is *to consider social facts as things*."[36] And many political scientists almost instinctively adhere to that principle. Nonetheless, the notion that much of what is studied is socially constructed raises some thorny epistemological issues.

Constructionism and Critical Theory

Most political scientists take reality pretty much as a given. That is, they posit that the objects they study—elections, wars, constitutions, government agencies—have an existence independent of observers and can be studied more or less objectively. But an alternative perspective, called the social construction of reality or **constructionism**,[37] casts doubt on this uncritical, perhaps blasé attitude. According to constructionism, humans do not simply discover knowledge of the real world through neutral processes, such as experimentation or unbiased observation; rather, they *create* the reality they analyze. This position is perhaps another way of saying, "Facts do not speak for themselves but are always interpreted or constructed by humans in specific historical times and settings." This stance may be likened to Umpire 3, who you may recall says, "They [the phenomena under investigation] aren't anything until I call them," as though the very act of umpiring creates its own reality.

One version of this position admits that entities (for example, molecules, planets) exist separately from anyone's thoughts about them, but it also insists that much of what people take for granted as being "real" or "true" of the world is built from learning and interaction with others and does not have an existence apart from human thought.[38] Consider the term *Democratic Party.* Instead of having an independent, material existence like an electron or a strand of DNA, a political party exists only because citizens behave as if it did exist. This means that two individuals

[35] Colin McGinn, "Is Just Thinking Enough?" review of *Making the Social World: The Structure of Human Civilization,* by John R. Searle, *New York Review of Books,* November 11, 2010, 58. Available at http://www.nybooks.com/articles/archives/2010/nov/11/just-thinking-enough/

[36] Émile Durkheim, *The Rules of Sociological Method and Selected Texts on Sociology and Method,* ed. Steven Lukes (New York: The Free Press, 1982), 60. (Emphasis in original.)

[37] The term *constructionism* encompasses an enormous variety of philosophical perspectives, the description of which goes far beyond the purposes of this book. The seminal work that brought the ideas into sociology and from there into political science is Peter L. Berger and Thomas Luckmann, *The Social Construction of Reality* (New York: Doubleday, 1966). An excellent but challenging analysis of constructionism is Ian Hacking, *The Social Construction of What?* (Cambridge, Mass.: Harvard University Press, 1999). Equally important, members of this school have widely varying opinions about the place of empiricism in social research. Many constructivists feel their position is perfectly consistent with the scientific study of politics; others do not.

[38] See John R. Searle, *The Construction of Social Reality* (New York: Free Press, 1995).

who come from different social, historical, and cultural backgrounds may not comprehend and respond to the term in the same way. What is important in studying, say, individuals' responses to Democratic candidates is fathoming their personal beliefs and attitudes about the party.

Constructionist thinking now plays a strong role in international relations theory, where a concept such as *anarchy* is not considered a "given and immutable" cause of the behavior of states (for example, their desire for security through power politics). Rather, concepts like this one have to be understood in terms of what actors (individuals, states) make of them.[39]

The constructionist viewpoint, which comes in innumerable varieties, challenges the idea of an objective epistemology, or theory of knowledge. Such ideas, however, are of a deeply methodological nature and raise deep philosophical issues that go well beyond the task of describing the empirical methods used in the discipline.[40] We thus acknowledge that the scientific study of politics is controversial but nevertheless maintain that the procedures we describe in the chapters that follow are widely accepted and can in many circumstances lead to valuable understandings of political processes and behaviors. Moreover, they have greatly shaped the research agenda and teaching of the discipline, as can be seen by looking at the evolution of the field in the twentieth century.

The emergence and domination of the empirical perspective have also brought about renewed interest in normative philosophical questions of "what ought to be" rather than "what is."[41] Part of the discipline has become receptive to variations of **critical theory**, or the belief that a proper goal of social science is to critique and improve society (by making it more just and humane) rather than merely understand or explain what is going on. Critical theorists feel, in other words, that by simply analyzing a polity as it is amounts to a tacit endorsement of its institutions and the distribution of power. Contrary to the idea that science should be value-free, critical theorists argue that proposing and working for reforms are legitimate activities for the social sciences. They therefore analyze institutions, practices, ideologies, and beliefs not only for their surface characteristics but also for their "hidden meanings" and implications for behavior.

Take, for example, the statement "I'm just not interested in politics."[42] An empirical political scientist might take this simply as a cut-and-dried case of apathy. He or

[39] Wendt, "Anarchy Is What States Make of It."

[40] For an excellent collection of articles about the pros and cons of studying human behavior scientifically, see Michael Martin and Lee C. Anderson, eds., *Readings in the Philosophy of Social Science* (Cambridge, Mass.: MIT Press, 1996).

[41] Isaak, *Scope and Methods*, 45.

[42] This example is based on an article by Isaac D. Balbus, "The Concept of Interest in Pluralist and Marxian Analysis," *Politics & Society* 1, no. 2 (1971): 151–77.

she might then look for variables (e.g., age, gender, ethnicity) associated with "not interested" responses on questionnaires. A critical theorist, by contrast, might ask, "Does this person really have *no* interest in current events? After all, isn't everyone affected by most political outcomes, like decisions about taxes, war and peace, and the environment, and thus in fact *have* an interest in politics? So perhaps we have a case of, say, 'false consciousness,' and it is crucial to uncover the reasons for lack of awareness of one's 'real' stake in politics. Is the indifference a matter of choice, or does it stem from the (adverse) effects of the educational system, the mass media, modern campaigning, or some other source?"

Here is another case. An important challenge to research in political science (as well as in other social science disciplines, such as sociology) has come from feminist scholars. Among the criticisms raised is that "the nature of political action and the scope of political research have been defined in ways that, in particular, exclude *women as women* [emphasis added] from politics."[43] Accordingly, "what a feminist political science must do is develop a new vocabulary of politics so that it can express the specific and different ways in which women have wielded power, been in authority, practiced citizenship, and understood freedom."[44] Even short of arguing that political science concepts and theories have been developed from a male-only perspective, it is all too easy to point to examples of gender bias in political science research. Examples of such bias include failing to focus on policy issues of importance to women, assuming that findings apply to everyone when the population studied was predominantly male, and using biased wording in survey questions.[45]

A related complaint is that political science in the past ignored the needs, interests, and views of the poor, the lower class, and the powerless and served mainly to reinforce the belief that existing institutions were as good as they could be. Those who agree with this complaint are called "critical theorists." Concerns about the proper scope and direction of political science have not abated, although nearly all researchers and teachers accept the need to balance the scientific approach with consideration of practical problems and moral issues.[46]

Let's wrap up our discussion so far before returning to the all-important question: What difference does all this philosophizing make? Table 2-1 lists some of the key differences between what we have been calling the empirical and nonempirical schools.

[43] Kathleen B. Jones and Anna G. Jonasdottir, "Introduction: Gender as an Analytic Category in Political Science," in *The Political Interests of Gender,* ed. Kathleen B. Jones and Anna G. Jonasdottir (Beverly Hills, Calif.: Sage, 1988), 2.

[44] Kathleen B. Jones, "Towards the Revision of Politics," in *The Political Interests of Gender,* ed. Kathleen B. Jones and Anna G. Jonasdottir (Beverly Hills, Calif.: Sage, 1988), 25.

[45] Margrit Eichler, *Nonsexist Research Methods: A Practical Guide* (Boston: Allen and Unwin, 1987).

[46] See the articles comprising "Political Science and Political Philosophy: A Symposium," *PS: Political Science and Politics* 33, no. 2 (2000): 189–97.

TABLE 2-1 **Methodological Perspectives in Political Science**

	Non-empirical	Empirical
Goals	To understand behavior To interpret actions	Causal explanations and predictions of individual and institutional behaviors General theory and laws Information of practical use "Value-free" knowledge
Assumptions	Social facts (at least) are "constructed." Institutions are social creations. Objective observation is not generally possible because our very senses are affected by culturally defined and imposed prior beliefs. Totally value-free research is impossible.	Realism (appearance and reality are the same). Independent, objective observation is possible. Behavior and, implicitly, institutions exhibit regularities. Claims about the real world must be verified. Attitudes (values, biases, beliefs) must not affect observation and analysis. There are no causeless effects.
Basic toolkit	Qualitative	Quantitative
Methods	Qualitative analysis (e.g., ethnography, content and document analysis, study of discourse) Case studies and comparisons	Case studies and comparisons Experiments and field experiments Mathematical models Surveys Statistical analysis of data Simulations
Objections	Observation is impressionistic, subjective, and nonsystematic. Knowledge is "nontransmissible." Findings are tainted by the investigator's values and biases.	Takes "politics out of political science." Concentration on formalism, quantitative measurement, and mathematical analysis leads to trivial and practically meaningless results.
Alleged biases	Conclusions are affected by political and social ideologies.	Inherently favors the status quo and existing power structures.

Source: This table is based partly on tables in Colin Hay, *Political Analysis: A Critical Introduction* (New York: Palgrave, 2002), chap. 1.

Conclusion

In this chapter we described the characteristics of scientific knowledge and the scientific method. We presented reasons why political scientists are attempting to become more scientific in their research and discussed some of the difficulties associated with empirical political science. We also touched on questions about the value of the scientific approach to the study of politics. Despite these difficulties and uncertainties, the empirical approach is widely embraced, and students of politics need to be familiar with it. In chapter 3 we begin to examine how to develop a strategy for investigating a general topic or question about some political phenomenon scientifically.

TERMS INTRODUCED

Actions. Human behavior done for a reason.

Constructionism. An approach to knowledge that asserts humans actually construct—through their social interactions and cultural and historical practices—many of the facts they take for granted as having an independent, objective, or material reality.

Critical theory. The philosophical stance that disciplines such as political science should assess society critically and seek to improve it, not merely study it objectively.

Cumulative. Characteristic of scientific knowledge; new substantive findings and research techniques are built upon those of previous studies.

Empiricism. Relying on observation to verify propositions.

Explanatory. Characteristic of scientific knowledge; signifying that a conclusion can be derived from a set of general propositions and specific initial considerations; providing a systematic, empirically verified understanding of why a phenomenon occurs as it does.

Falsifiability. A property of a statement or hypothesis such that it can (in principle, at least) be rejected in the face of contravening evidence.

Interpretation. Philosophical approach to the study of human behavior that claims that one must understand the way individuals see their world in order to truly understand their behavior or actions; philosophical objection to the empirical approach to political science.

Nonnormative knowledge. Knowledge concerned not with evaluation or prescription but with factual or objective determinations.

Normative knowledge. Knowledge that is evaluative, value-laden, and concerned with prescribing what ought to be.

Parsimony. The principle that among explanations or theories with equal degrees of confirmation, the simplest—the one based on the fewest assumptions and explanatory factors—is to be preferred; sometimes known as Ockham's razor.

Social facts. Values and institutions that have a subjective existence in the minds of people living in a particular culture.

Theory. A statement or series of related statements that organize, explain, and predict phenomena.

Transmissible. Characteristic of scientific knowledge; indicates that the methods used in making scientific discoveries are made explicit so that others can analyze and replicate findings.

Verification. The process of confirming or establishing a statement with evidence.

SUGGESTED READINGS

Box-Steffensmier, Janet, Henry Brady, and David Collier. *The Oxford Handbook of Political Methodology.* New York: Oxford University Press, 2008.

Brady, Henry E., and David Collier, eds. *Rethinking Social Inquiry: Diverse Tools, Shared Standards.* Lanham, Md.: Rowman and Littlefield, 2004.

Elster, Jon. *Nuts and Bolts for the Social Sciences.* Cambridge, UK: Cambridge University Press, 1990.

Hay, Colin. *Political Analysis: A Critical Introduction.* New York: Palgrave, 2002.

Hindmoor, Andrew. *Rational Choice.* New York: Palgrave Macmillan, 2006.

King, Gary, Robert O. Keohane, and Sidney Verba. *Designing Social Inquiry: Scientific Inference in Qualitative Research.* Princeton, N.J.: Princeton University Press, 1994.

Kuhn, Thomas. *The Structure of Scientific Revolutions.* 2nd ed. Chicago, Ill.: University of Chicago Press, 1971.

Nielsen, Joyce McCarl, ed. *Feminist Research Methods: Exemplary Readings in the Social Sciences.* Boulder, Colo.: Westview, 1990.

Rosenberg, Alexander. *The Philosophy of Social Science.* 3rd ed. Boulder, Colo.: Westview, 2007.

Silver, Brian L. "I Believe." Chap. 2 in *The Ascent of Science.* New York: Oxford University Press, 1998.

CHAPTER 3

Beginning the Research Process:

Identifying a Research Topic, Developing Research Questions, and Reviewing the Literature

CHAPTER OBJECTIVES

3.1 Explain the purpose of specifying a research question.

3.2 Identify different sources of ideas for research topics.

3.3 Summarize the reasons why conducting a literature review is helpful.

3.4 Describe the steps in collecting sources for a literature review.

3.5 Discuss how to approach writing a literature review.

3.6 Relate the basic organizational structure of a literature review.

MANY STUDENTS FIND CHOOSING AN APPROPRIATE research topic to be a challenging part of the research process. In this chapter, therefore, we discuss general attributes of promising research topics, suggest some methods for discovering interesting topics and research questions, and provide guidelines for conducting a systematic review of the literature on a topic and tips on writing a literature review—an important component of all academic articles and research reports.

Specifying the Research Question

One of the most important purposes of research is to answer questions about social phenomena. The research projects summarized in chapter 1, for example, attempt to answer questions about some important political attitudes or behaviors: Why is wealth distributed more equally among the population in some countries than in others? Why do some people vote in elections while others do not? Do Supreme Court decisions affect people's opinions on issues and people's support of the Supreme Court? Is the protection of some types of human rights linked to the protection of other types? Does economic growth lead to more democratic institutions and practices and fewer human rights abuses? Under what circumstances are people most likely to support US involvement in foreign affairs? What factors limit the public's tolerance for war casualties? Does negative campaign advertising have any impact on the electorate? How do interest groups influence the extent to which members of Congress engage in oversight of agency decisions? Do women exhibit the same level of political ambition as men? If not, why? In each case, the researchers identified a political phenomenon that interested them and tried to answer questions about that phenomenon.

The phenomena investigated by political scientists are diverse and are limited only by whether they are significant (that is, would advance our understanding of politics and government), observable, and political. Political scientists attempt to answer questions about the political behavior of individuals (voters, citizens, residents of a particular area, Supreme Court justices, members of Congress, presidents), groups (political parties, interest groups, labor unions, international organizations), institutions (state legislatures, city councils, bureaucracies, district courts), and political jurisdictions (cities, states, nations).

Most students, when confronting a research project for the first time, do not have a well-formulated research question as their starting point. Some will start by saying, "I'm interested in X," where X may be the Supreme Court, media coverage of a policy issue such as climate change, public attitudes about Congress, or some other political phenomenon. Others may not have any specific interest or topic in mind at all. Thus, the first major task in a research effort often is to find a topic and to translate a general interest in a topic into a manageable research question or series of questions or propositions. Framing an engaging and appropriate research question will get a research project off to a good start by defining, and limiting, the scope of the investigation and determining what information has to be collected to answer the question. A poorly specified question inevitably leads to wasted time and energy.

Any of the following questions would probably lead to a politically significant and informative research project:

- Why is voter turnout for local elections higher in some cities than in others?
- Why does the amount spent per pupil by school districts vary (within a state or among states)?

- Do small nations sign more multilateral treaties than large nations?
- Why did some members of Congress vote for the Restoring Financial Stability Act of 2010, whereas others opposed it?
- Does the legislative output of legislatures change after term limits have gone into effect?
- Why do some nations (or states) have cap-and-trade programs for carbon dioxide emissions while others do not?
- Do independents have more moderate views on major political issues than those who identify themselves as strong partisans?

A research project will get off on the wrong foot if the question that shapes it fails to address a political phenomenon, is unduly concerned with discrete facts, or is focused on reaching normative conclusions. Although the definition of *political phenomenon* is vague, it does not include the study of *all* human characteristics or behaviors.

Research questions, if they dwell on discrete or narrow factual issues, may limit the significance of a research project. Although important, facts alone are not enough to yield scientific explanations. What is missing is a **relationship**—that is, the association, dependence, or covariance of the values of one variable with the values of another. Researchers are generally interested in how to advance and test generalizations relating one phenomenon to another. In the absence of such generalizations, factual knowledge of the type called for by the following research questions will be fundamentally limited in scope:

- How many seats in the most recent state legislative elections in your state were uncontested (had only one contestant)?
- How many members of Congress had favorable environmental voting records in the last session of Congress?
- How many trade disputes have been referred to the World Trade Organization (WTO) for resolution in the past five years?
- How many local governments have voted on proposals to ban hydraulic fracturing? How many such proposals have been adopted?
- How much money was spent on campaign advertising by independent groups in your state in the last election cycle?

Factual information, however, may lead a researcher to ask "why" questions. For example, if a researcher has information about the number of uncontested seats and notes that this number varies substantially from state to state, the research question, "Why are legislative elections competitive in some states and not in others?" forms the basis for an interesting research project. Alternatively, if one had data from just one state, one could investigate the question, "Why do some districts have competitive elections and not others?" This would involve identifying characteristics of districts and elections that might explain the difference.

Or someone might notice that the number of trade disputes referred to the WTO has varied from year to year. What explains this fluctuation? When collecting data on the number of disputes, the researcher might notice that the complaints originate in many different countries. It would be interesting, then, to find out how the disputes are resolved. Is there any pattern to their resolution in regard to which countries benefit or the principles and arguments underlying the decisions? Why?

Similarly, the environmental voting records of members of Congress differ. Why? Is political party a likely explanation? Is ideology? Or is some other factor responsible? Furthermore, there aren't many communities that have voted on the question whether or not to ban "fracking." Why is this so? Do local residents generally support the practice? Do all states in which fracking takes place allow local governments to vote on such a question? If not, why not?

Sometimes, important research contributions come from descriptive or factual research because the factual information being sought is difficult to obtain or, as we discuss in chapter 5, disagreement exists over which information or facts should be used to measure a concept. In this situation, a research effort will entail showing how different ways of measuring a concept have important consequences for establishing the facts. For example, how income inequality should be measured is certainly an important aspect of research on that topic.

Questions calling for normative conclusions also are inconsistent with the research methods discussed in this book. (Refer to chapter 2 for the distinction between normative and empirical statements.) For example, questions such as "Should the United States adopt a period of compulsory military service for all young adults?" or "Should a new federal agency be placed within the Executive Office of the President or should it be created as an independent agency?" or "Should states give tax breaks to new businesses willing to locate within their borders?" are important and suitable for the attention of political scientists (indeed, for any citizen), but they are inappropriate as framed here. As written, they ask for a normative response, seeking an indication of what is good or of what should be done. Although scientific knowledge may be helpful in answering questions like these, it cannot provide the answers without regard for an individual's personal values or preferences. Ultimately, the answers to these questions involve what someone likes or dislikes, values or rejects.

Normative questions, however, may lead you to develop an empirical research question. For example, a student of one of the authors felt that Pennsylvania's method of selecting judges using partisan elections was not a good way to choose judges. To contribute to an informed discussion of this issue, she collected data on the amount of money raised and spent by judicial candidates, the amount of money spent per vote cast in judicial races compared with that spent in other state elections, and the voter turnout rate in judicial races as compared with other races. This information spoke to some of the arguments raised against partisan judicial elections, but she discovered that it was very difficult to collect empirical evidence to answer the interesting question of whether reliance on campaign contributions jeopardized the independence and impartiality of judges.

Sources of Ideas for Research Topics

Potential research topics about politics come from many sources. These sources may be classified as personal, nonscholarly, or scholarly. Personal sources include your own life experiences and political activities and those of your family and friends, as well as class readings, lectures, and discussions.

You can also look to nonscholarly sources for research topics, including print, broadcast, and Internet sources. Becoming aware of current or recent issues in public affairs will help you develop interesting research topics. You can start by reading a daily newspaper or issues of popular magazines that deal with government policies and politics. The Web site accompanying this book (http://psrm.cqpress.com/) offers many possibilities and lists of other Web sites. The best print sources include national newspapers and magazines featuring in-depth political coverage. First, consider reading major urban daily newspapers like the *New York Times* and the *Washington Post.* Daily newspapers provide the most up-to-date printed political news and discussions and often draw attention to recently issued government reports. For example, many national media outlets picked up on a 2014 *National Vital Statistics* report on national and state patterns of teen births between 1940 and 2013.[1] The report presented data showing that teen birth rates in the United States had declined significantly, yet also indicated that there was considerable variation among the states in terms of teen birth rates and the speed at which their rates had dropped. Students in one of our own methods classes conducted research in which they proposed explanations for the variation in state rates and tested them with data they collected. In addition to daily news sources, look at weekly magazines like *Atlantic, Harper's,* the *Economist,* the *American Prospect, National Review,* the *New Republic,* the *New Yorker,* and the *Weekly Standard.* Most of these weeklies have a decidedly partisan leaning (either conservative or liberal, Republican or Democrat), but—and this is a key point—they contain serious discussions of domestic and foreign government and politics and are wonderful sources of ideas and claims to investigate.[2] Each of these sources also features online material, much of which is free.

[1] Stephanie J. Ventura, Brady E. Hamilton, and T. J. Mathews, "National and State Patterns of Teen Births in the United States, 1940–2013," *National Vital Statistics Reports* 63, no. 4 (2014). Accessed January 26, 2015. Available at http://www.cdc.gov/nchs/data/nvsr/nvsr63/nvsr63_04.pdf

[2] The *Political Science Research Methods* CD contains several text documents that illustrate this point and allow the reader to extract empirical and testable claims from verbal arguments.

An underappreciated source of potential research topics within these printed sources is the editorial and letters-to-the-editor pages. Although these pieces express opinions, the writers often support them with what they claim are empirical facts. Consider the statement that suggests that strict state gun control laws do not reduce homicides: "Utah has the nation's most permissive gun laws, according to the Brady Campaign to Prevent Gun Violence, but it has one of the lowest murder rates in the country. California, with the strictest laws, has a homicide rate higher than the national average."[3] (Maybe California has gun laws because it has a crime problem; it's not likely the presence of such laws cause an increase/excess of gun violence or that such laws are irrelevant. Maybe they prevent even more murders. These ideas could be investigated with more data.)

Broadcast news sources can also inspire topical research projects. The best radio and television programs for this purpose are those that include long segments dedicated to political news, discussion, and debate. Radio programs with a civic or political focus featuring a variety of topics include National Public Radio's *Morning Edition* and *All Things Considered.* Your local public radio station may have a program devoted to local and regional public affairs such as *Radio Times,* a daily two-hour program emanating from Philadelphia's public radio station, WHYY. Television shows such as NBC's *Meet the Press,* CBS's *Face the Nation,* ABC's *This Week,* Fox's *News Sunday,* PBS's *Charlie Rose Show* and the *News Hour with Jim Lehrer,* and investigative journalism programs like CBS's *60 Minutes* tend to feature long interviews with political actors. Numerous highly partisan or ideological political talk shows do not hesitate to make assertions about political matters that can be put to the empirical test.

Internet sources can include the print and broadcast sources discussed above, found through the publications' and broadcasts' sites on the Web. In addition to offering the same content that is printed or broadcast, many print and broadcast sources feature exclusive Internet material. An example is the *Washington Post*'s *White House Watch,* a daily blog focusing on the presidency. Other Internet sources include government, university, or organization Web sites; Web sites created by individuals; and political blogs. Blogs like *Daily Kos* or *InstaPundit* have become fixtures in the national political debate, raising topics or uncovering evidence that the traditional news media have not. Blogs, much like talk radio or magazines, often feature political discussion and debate from a particular ideological or partisan perspective.

Although personal and nonscholarly sources are good places to find potential research topics, surveying the scholarly literature will help you identify a topic

[3] Steve Chapman, "Restricting 2nd Amendment Isn't the Answer," *Real Clear Politics,* January 13, 2011. Available at http://www.realclearpolitics.com/articles/2011/01/13/restricting_2nd_amendment_isnt_the_answer.html

HELPFUL HINTS

How to Come Up with a Research Topic

- Pose a "how many" question. Where possible, collect data for more than one time (e.g., year, election) or for more than one case (e.g., more than one city, state, nation, primary election). Do any patterns emerge? What might explain these patterns?
- Is it difficult to find information to answer a question? Why? Could you make a meaningful contribution by collecting appropriate data?
- Do you think that the ways in which other researchers have measured the phenomena or concepts that interest you are adequate? Are there any validity or reliability problems with the measures? (Measurement validity and reliability are discussed in chapter 5.)
- Find an assertion or statement in the popular press or a conclusion in a research article that you believe to be incorrect. Look for empirical evidence so that you can assess the statement or examine the evidence used by the author to see if any mistakes were made that could have affected the conclusion.
- Find two studies that reach conflicting conclusions. Try to explain or reconcile the conflict. Test conflicting explanations by applying them to different cases or data.
- Take a theory or general explanation for certain political behaviors and apply it to a new situation.

Note: We wish to thank one of our anonymous reviewers for suggesting that we include tips for coming up with paper topics and for suggesting some of the tips listed here.

Check out more Helpful Hints at **edge.sagepub.com/johnson8e**

relevant to the discipline. The scholarly literature includes books and articles written by political scientists and other academics or political practitioners. Such literature establishes which topics and questions are important to political scientists. Simply perusing the list of article titles of several issues of a journal can lead to ideas for a topic. Here is a short list of some of the major political science publications, many (if not all) of which are available online:

American Journal of Political Science—Broad coverage of political science and public administration.

American Political Science Review—The official journal of the American Political Science Association.

British Journal of Political Science—Although emphasizing a comparative perspective, this publication contains important research on American political institutions and behaviors.

Comparative Politics—Begin here when looking for scholarly studies on all aspects of cross-national politics and government.

International Organization—Contains important articles on international relations. One of the leading journals in the field.

Journal of Conflict Resolution—A widely cited journal with articles on, among other topics, international relations, war and peace, and individual attitudes and behaviors. Authors use a variety of methods and research designs.

Journal of Politics—Broad coverage of political science and public administration.

Legislative Studies Quarterly—Articles about legislative organization and functioning and electoral behaviors.

Political Analysis—For students with a serious interest in methods and statistics. Articles frequently contain important substantive results.

Political Research Quarterly—Broad coverage of political science and public administration.

Polity—Articles on American politics, comparative politics, international relations, and political philosophy.

Social Science Quarterly—Articles on a wide range of topics in the social sciences.

World Politics—Analytical and theoretical articles, review articles, and research notes in international relations, comparative politics, political theory, foreign policy, and modernization.

Still another source of ideas for research papers is a textbook used in substantive courses, such as American politics, comparative politics, or international relations. These works can be particularly valuable for pointing out controversies within a field. For example, as the discussion of judicial behavior in chapter 1 of this textbook illustrated, political scientists argue about what underlies judges' decisions, political ideology, or adherence to legal precedent and principles. You might do a case study of a particular justice to see which side this person's rulings seem to support.

To guide you further in finding topics and searching for appropriate sources, this book's companion Web site lists additional professional journals as well as indexes and bibliographies, data banks, guides to political resources, and the like. A reference librarian will undoubtedly be able to provide additional information and guidance on particular library sources available.

So far, we have talked about using a variety of sources, including the scholarly literature, to help you identify a research topic of interest to you in a general sense. We haven't yet indicated how you might search the literature (both scholarly and nonscholarly) once you have at least a general interest in a topic. Before we show you how to conduct a search of the literature, however, we want to talk about why every serious research project conducts what is called a **literature review** and why scholarly articles and books contain a section or a chapter in which the literature related to the topic is discussed.

Why Conduct a Literature Review?

Most research topics are initially much too broad to be manageable. It would be virtually impossible to write something new on "international terrorism" or even "the causes of terrorism in the Middle East" without first knowing a great deal about the subject. Good research, therefore, involves reviewing previous work on the topic to motivate and sharpen a research question. Among the many reasons for doing so are (1) to see what has and has not been investigated, (2) to develop general explanations for observed variations in a behavior or a phenomenon, (3) to identify potential relationships between concepts and to identify researchable hypotheses, (4) to learn how others have defined and measured key concepts, (5) to identify data sources that other researchers have used, (6) to develop alternative research designs, and (7) to discover how a research project is related to the work of others. Let us examine some of these reasons more closely.

Often, someone new to empirical research will start out by expressing only a general interest in a topic, such as terrorism or the effects of campaign advertising or public opinion and international relations, but the specific research question has yet to be formulated (for example, "What kinds of people become terrorists?" or "Do negative televised campaign advertisements sway voters?" or "Does the public support isolationism or internationalism?"). A review of previous research can help you sharpen a topic by identifying research questions that others have asked.

Alternatively, you may start with an overly specific research question such as "Do married people have different views on abortion policy than those who are single?" Reading the literature related to public opinion on abortion likely will reveal that your specific research question is one of many aimed at answering the more general research question: What are the characteristics or attributes of people who oppose abortion, and do they differ from those of supporters? This latter research question constitutes a topic, whereas the former is likely to be too narrow to sustain a research paper.

After reading the published work in an area, you may decide that previous reports do not adequately answer the question. Thus, you may design a research project to answer an old question in a new way. An investigation may replicate a study

HELPFUL HINTS

Differentiating Scholarly from Nonscholarly Literature

You can differentiate scholarly works from nonscholarly ones by looking for a few characteristics. Most important, professional articles and books published in political science or other disciplines will often go through a peer-review process. The most common peer-review standard is that a journal or book editor sends an article or book manuscript submitted for publication to one or more scholars with expertise in the topical area of the article. The review is performed in a blind fashion; that is, the reviewers are not told the author's name to ensure that reviewers assess only the quality of the work. The editor relies on the peer reviewers' comments to suggest revisions of the work and assess whether or not the work makes a sufficient contribution to the literature to deserve publication. The peer-review process helps ensure that the work published in scholarly journals and books is of the best possible quality and of the most value to the discipline. It also assures the reader that, although there still may be mistakes or invalid or unreliable claims, the article or book has been vetted by one or more experts on the topic.

Alternatively, some scholarly journals and books are reviewed only by the editorial staff. Although this method provides a check on the quality of the work, it is usually not as rigorous as a blind peer review. The type of review a journal or book publisher uses will typically be explained in the journal or on the journal or publisher's Web site.

In addition to a peer-review process, some other indicators can differentiate scholarly from nonscholarly work. Scholarly articles and books are usually written by academics, journalists, political actors, or other political practitioners, so looking for a description of the authors is the place to start. Scholarly books are published by both university presses and commercial presses. If you are still unsure about whether or not a particular work is scholarly, consult with a reference librarian or your instructor.

Check out more Helpful Hints at **edge.sagepub.com/johnson8e**

to confirm or challenge a hypothesis or expand our understanding of a concept. Replication is one of the cornerstones of scientific work. By testing the same hypothesis in different ways or confirming the results from previous research using the same data and methods, we increase our confidence that the results are correct. Replication can therefore help build consensus or identify topics that require further work.

At other times, research may begin with a hypothesis or with a desire to develop an explanation for a relationship that has already been observed. Here, a literature review may reveal reports of similar observations made by others and may also help you develop general explanations for the relationship by identifying theories that explain the phenomenon of interest. Your research will be more valuable if you can provide a general explanation of the observed or hypothesized relationship rather than simply a report of the empirical verification of a relationship.

In addition to seeking theories that support the plausibility and increase the significance of a hypothesis, you should be alert for competing or alternative hypotheses. You may start with a hypothesis specifying a simple relationship between two variables. Since it is uncommon for one political phenomenon to be related to or caused by just one other factor or variable, it is important to look for other possible causes or correlates of the dependent variable. Data collection should include measurement of these other relevant variables so that, in subsequent data analysis, you may rule out competing explanations or at least indicate more clearly the nature of the relationship between the variables in the original hypothesis.

Collecting Sources for a Literature Review

After selecting a research topic using the sources described above, you must begin collecting sources for use in writing a literature review. Although personal and nonscholarly sources can be quite helpful in selecting a research topic, and a literature review can encompass virtually anything published on your topic, we strongly encourage you to become familiar with the scholarly literature. Relying on scholarly sources rather than nonscholarly ones will improve the quality of a literature review. In addition, as a practical concern, many instructors may not accept or give much credit for citations from nonscholarly sources unless their content constitutes part of your topic. After all, a literature review is supposed to establish the knowledge about a topic that has been attained and communicated according to professional or scientific principles.

Students commonly ask, "How many sources must I find to write my literature review?" The answer, unfortunately, is not straightforward. How many books and articles to include in a literature review depends on the purpose and scope of the project, as well as available resources. If your project is focused largely on reporting the work of others, you will probably need to include more sources than if your project is focused mostly on your own analysis. Furthermore, a more complex topic, or a topic with a larger literature, may require a more in-depth literature review than will a more straightforward topic or one with a smaller literature. Finally, consider how much time and effort you are willing to dedicate to collecting sources. Although we cannot provide a simple answer to the question of how many sources are necessary, we can explain how available time and effort could be best directed and used most efficiently.

Identifying the Relevant Scholarly Literature

It would be impossible for anyone to identify, let alone read and/or write about, every book or article with relevance to any particular research project. With that caveat in mind, you can think of the first step in collecting sources—identifying the relevant literature—as limiting the search to only those books and articles with the most direct relevance to the research topic of interest. You can begin to narrow the field of potential sources in many ways. The first step is to search comprehensive **electronic databases**, such as Web of Science or Google Scholar, or other databases that include links to full text articles, such as JSTOR. These databases allow you to quickly locate a large number of articles and possibly books and published conference proceedings related to your topic.

HELPFUL HINTS

Pyramid Citations

Each time you find what appears to be a useful source, look at its list of notes and references. One article, for example, may cite two more potentially useful papers. Each of these, in turn, may point to two or more additional ones, and so on. Even if you start with a small list, you can quickly assemble a huge list of sources. Moreover, you increase your chances of covering all the relevant literature.

Check out more Helpful Hints at **edge.sagepub.com/johnson8e**

Web of Science is a particularly useful starting point for building a literature review because you can

- search the Social Sciences Citation Index database of social science journal articles, books, and published conference proceedings generally, using a keyword search;
- search for articles written by a particular social scientist; or, perhaps most important for starting a literature review,
- search for all of the articles in the database that cite an article you know to be of interest and for articles that subsequently cited those articles.

Two quick examples highlight the value of these searches. First, suppose you are interested in understanding judicial behavior.

1. By typing the phrase *judicial behavior* into the "basic search" field, limiting the period to 1970–2015, and searching the Social Sciences Citation Index, the Conference Proceedings Index—Social Science & Humanities, and the Book Citation Index—Social Science & Humanities, we found 885 sources. This is far too many to sort through.
2. You might want to refine your search further by limiting subject areas to law and political science, sources just to articles and to the United States (a "Countries/Territories" option with a drop-down menu allows you to choose a country). Doing this reduced the number of items to 377. This is still a lot, but after reading through article abstracts in this larger topic, you might narrow the search to a particular kind of judicial behavior. For example, say that, after reading a few abstracts and articles, you found you were interested in the debate over judicial activism.
3. By entering the phrase *judicial activism* into the "search within these results" search field, we narrowed the search to only nine articles. This is a manageable number of articles to examine.
4. You may find that not all of the articles are actually relevant to your topic, but if you find one article that is of direct relevance, you can use it to find more like it using the approach we describe next.

A second way to use the Web of Science is to begin with a single article instead of searching for topics.

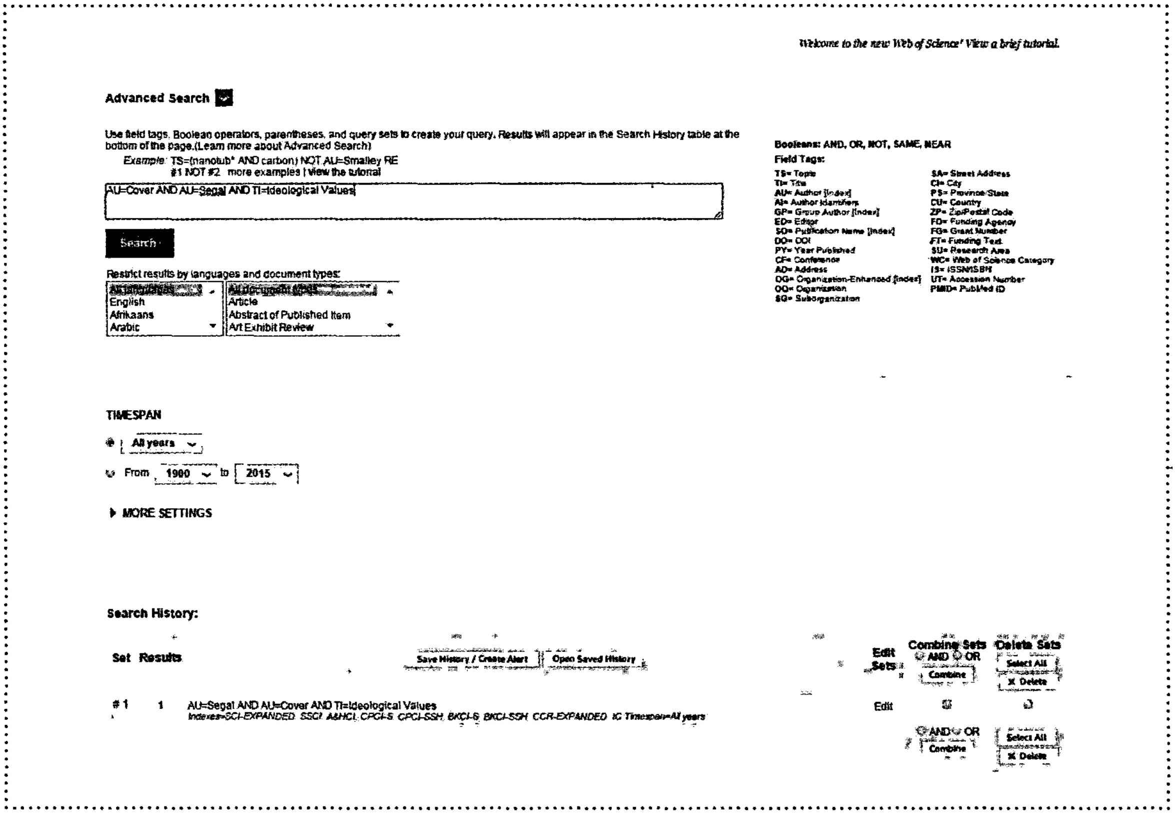
1. Suppose that at the beginning of your search, you decided to look for articles related to an article you had already found—from your course syllabus, for example. Assume that while reading Jeffrey A. Segal and Albert D. Cover's "Ideological Values and the Votes of U.S. Supreme Court Justices" (a brief discussion of the article is found in chapter 1), you found

that the topic interested you and thought you might like to find out what else was written on it.

2. With this single article, you could use the Web of Science to quickly find other work in the literature investigating similar research questions. For example, by using a basic search and selecting the Social Sciences Citation Index and the default of "all years," you could find Segal and Cover's article by searching for the authors' names and part of the article's title, as shown in figure 3-1. (Do not bother to restrict your search to the English language and to just articles, even though you know both of these items to be true in the case of the item for which you are searching. The authors did this and the search came up empty. No search engine is perfect, and casting as wide a net as is reasonable can avoid searches that for some reason fail to find relevant items.)
3. Click on the article title to find a wealth of information about the article. Figure 3-2 shows the full citation, the number of articles the article cited, and the number of articles that subsequently cited Segal and Cover's article.
4. Segal and Cover cited thirty-four references in their article; by clicking on "Cited References: 34," you will find all thirty-four references with electronic links to those references included in the Web of Science database. This feature makes it easy to review the base of knowledge that was in place before Segal and Cover's article.
5. As of this writing, Web of Science has identified 301 articles in its database that have cited Segal and Cover's article. By clicking on "Times Cited: 301," you can find a link to each of these articles. This is particularly important because once you find an essential reference like Segal and Cover's article—an article that most work in this literature includes as a reference—you can easily identify a large number of articles for your own literature review. In this example, we located 335 references to relevant books and articles in a matter of minutes.
6. If you click on "Times Cited," a list of all of the works that cited the Segal and Cover article will appear (see figure 3-3). You can work through this list and keep the ones that appear relevant judging from the title or clicking on the abstract where available. By clicking on the box to the left of the article number and then clicking on "Add to Marked List" before moving on to a new page of the results, you can collect just the articles you want to keep.
7. Finally, you can click on "Times Cited" for any articles in your marked list to see if any new and relevant sources can be identified.

The larger lesson from this example is that once you find a relevant article, you can sharpen the direction of your search for relevant literature by examining the

FIGURE 3-1 Basic Search on Web of Science Database



Welcome to the new Web of Science! View a brief tutorial.

Advanced Search

Use field tags, Boolean operators, parentheses, and query sets to create your query. Results will appear in the Search History table at the bottom of the page.(Learn more about Advanced Search)

Example: TS=(nanotub* AND carbon) NOT AU=Smalley RE
#1 NOT #2 more examples | View the tutorial

AU=Cover AND AU=Segal AND TI=Ideological Values

Search

Restrict results by languages and document types:

All languages
English
Afrikaans
Arabic

All document types
Article
Abstract of Published Item
Art Exhibit Review

Booleans: AND, OR, NOT, SAME, NEAR

Field Tags:

TS= Topic
TI= Title
AU= Author [Index]
AI= Author Identifiers
GP= Group Author [Index]
ED= Editor
SO= Publication Name [Index]
DO= DOI
PY= Year Published
CF= Conference
AD= Address
OG= Organization-Enhanced [Index]
OO= Organization
SG= Suborganization

SA= Street Address
CI= City
PS= Province/State
CU= Country
ZP= Zip/Postal Code
FO= Funding Agency
FG= Grant Number
FT= Funding Text
SU= Research Area
WC= Web of Science Category
IS= ISSN/ISBN
UT= Accession Number
PMID= PubMed ID

TIMESPAN

● All years
○ From 1900 to 2015

▸ MORE SETTINGS

Search History:

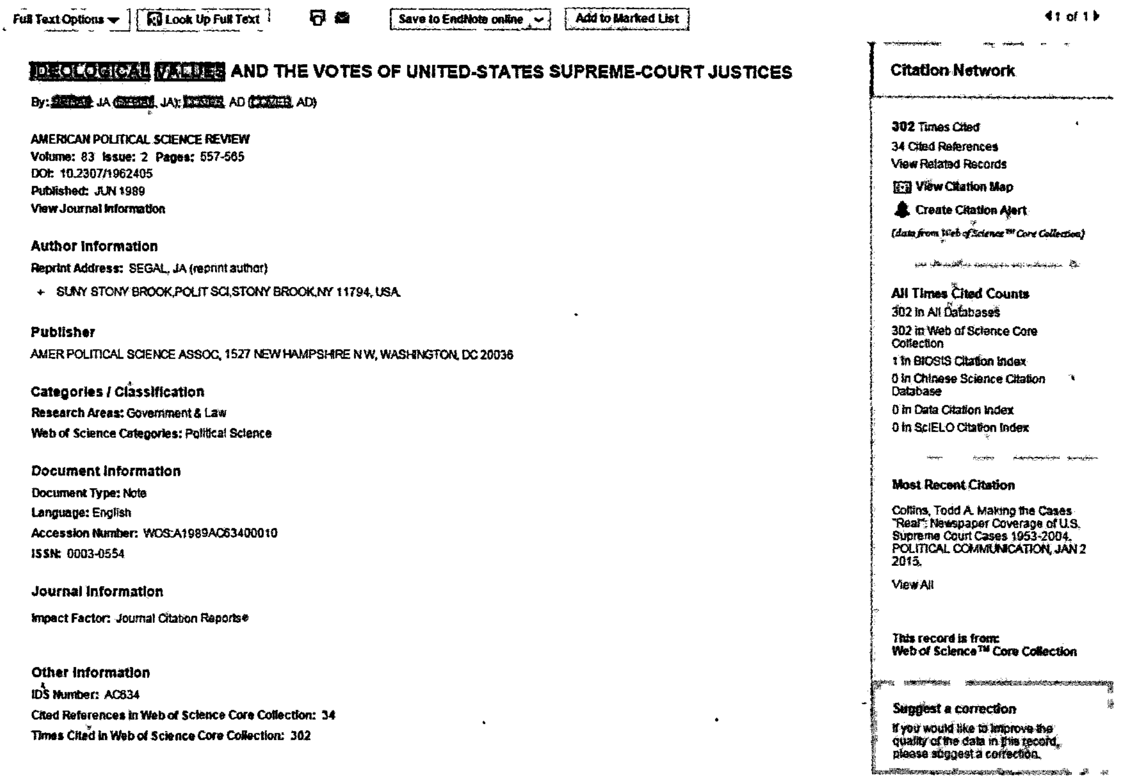
Set	Results		Edit Sets	Combine Sets	Delete Sets
# 1	1	AU=Segal AND AU=Cover AND TI=Ideological Values *Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, CCR-EXPANDED, IC Timespan=All years*	Edit	□	□

Source: Thomson Reuters, *Web of Science* database, www.isiknowledge.com.

literature review and works cited in that article. Since the article is directly relevant to the research topic of interest, the sources used in the article will likely be related as well. It is also quite likely that sources citing the relevant article are related to your topic. By building a list of sources in this fashion, you can save a great deal of time and effort as well as collect sources with a greater certainty that you will not overlook important work.

Remember, however, that even though both of the above example strategies will help you find relevant articles quickly, articles without much relevance may also come up in a search. Two articles that share a common **search term** do not necessarily have much related content. Nor does one article's citing another necessarily

FIGURE 3-2 Results of Clicking on Article Title on Web of Science Database



IDEOLOGICAL VALUES AND THE VOTES OF UNITED-STATES SUPREME-COURT JUSTICES

By: SEGAL JA (SEGAL, JA); COVER AD (COVER, AD)

AMERICAN POLITICAL SCIENCE REVIEW
Volume: 83 Issue: 2 Pages: 557-565
DOI: 10.2307/1962405
Published: JUN 1989
View Journal Information

Author Information
Reprint Address: SEGAL, JA (reprint author)
+ SUNY STONY BROOK,POLIT SCI,STONY BROOK,NY 11794, USA

Publisher
AMER POLITICAL SCIENCE ASSOC, 1527 NEW HAMPSHIRE N W, WASHINGTON, DC 20036

Categories / Classification
Research Areas: Government & Law
Web of Science Categories: Political Science

Document Information
Document Type: Note
Language: English
Accession Number: WOS:A1989AC63400010
ISSN: 0003-0554

Journal Information
Impact Factor: Journal Citation Reports®

Other Information
IDS Number: AC634
Cited References in Web of Science Core Collection: 34
Times Cited in Web of Science Core Collection: 302

Citation Network

302 Times Cited
34 Cited References
View Related Records
View Citation Map
Create Citation Alert
(data from Web of Science™ Core Collection)

All Times Cited Counts
302 in All Databases
302 in Web of Science Core Collection
1 in BIOSIS Citation Index
0 in Chinese Science Citation Database
0 in Data Citation Index
0 in SciELO Citation Index

Most Recent Citation
Collins, Todd A. Making the Cases "Real": Newspaper Coverage of U.S. Supreme Court Cases 1953-2004. POLITICAL COMMUNICATION, JAN 2 2015.
View All

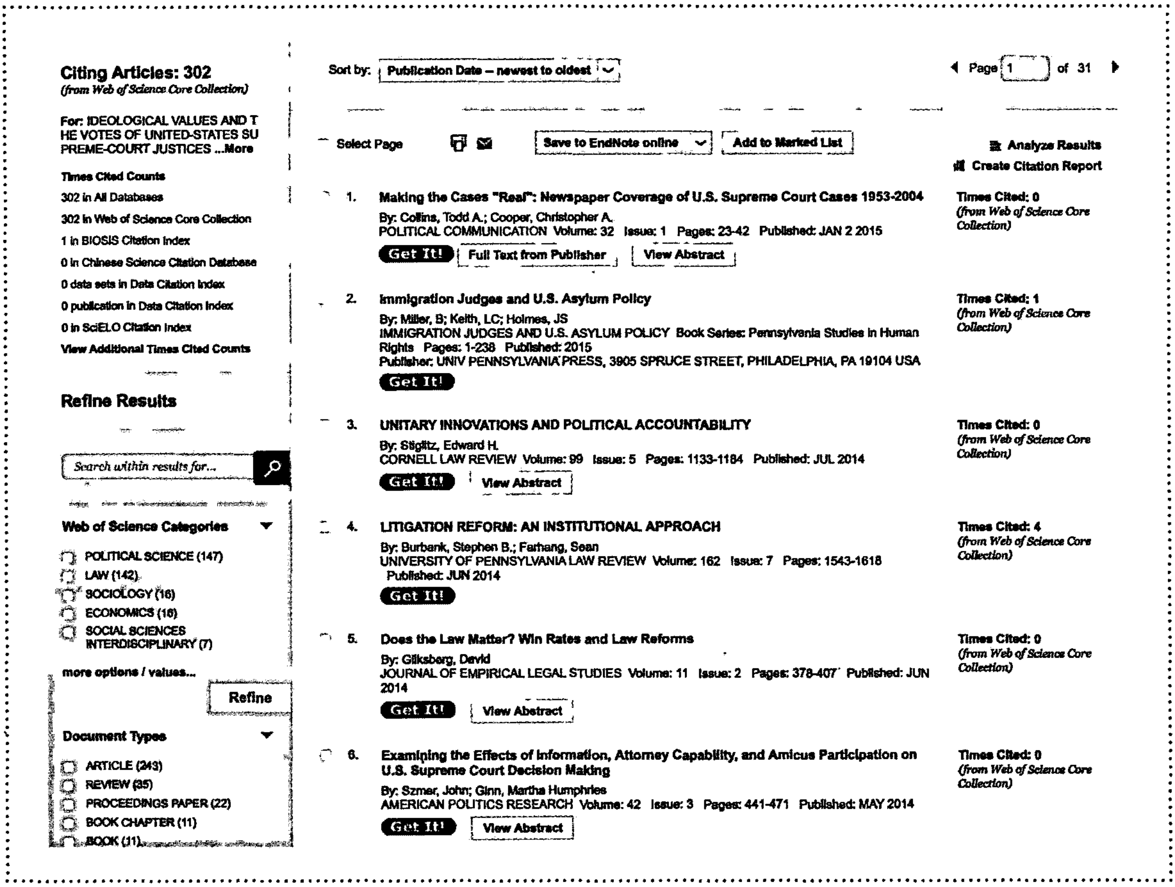
This record is from:
Web of Science™ Core Collection

Suggest a correction
If you would like to improve the quality of the data in this record, please suggest a correction.

Source: Thomson Reuters, *Web of Science* database, www.isiknowledge.com.

mean that the two articles investigate the same topic. Therefore, you should be prepared to review the lists of sources you identify and cull those that are not relevant to your topic. You could also search for articles on judicial behavior using a database like JSTOR, a comprehensive electronic archive of academic journals and publications. Although not every campus has access to it, and it does not include full-text articles from many important sources, JSTOR is widely available. When JSTOR and Web of Science do provide access to the full text of articles, you can save them to your computer or storage device, thus saving printing costs. Moreover, a description of how to search it illustrates guidelines for searching other databases.

FIGURE 3-3 Results of Clicking on "Times Cited" for Article on Web of Science Database



Citing Articles: 302
(from Web of Science Core Collection)

For: IDEOLOGICAL VALUES AND T HE VOTES OF UNITED-STATES SU PREME-COURT JUSTICES ...More

Times Cited Counts
302 in All Databases
302 in Web of Science Core Collection
1 in BIOSIS Citation Index
0 in Chinese Science Citation Database
0 data sets in Data Citation Index
0 publication in Data Citation Index
0 in SciELO Citation Index
View Additional Times Cited Counts

Refine Results

Web of Science Categories
- □ POLITICAL SCIENCE (147)
- □ LAW (142)
- □ SOCIOLOGY (16)
- □ ECONOMICS (16)
- □ SOCIAL SCIENCES INTERDISCIPLINARY (7)

more options / values...

Document Types
- □ ARTICLE (243)
- □ REVIEW (35)
- □ PROCEEDINGS PAPER (22)
- □ BOOK CHAPTER (11)
- □ BOOK (11)

1. **Making the Cases "Real": Newspaper Coverage of U.S. Supreme Court Cases 1953-2004**
By: Collins, Todd A.; Cooper, Christopher A.
POLITICAL COMMUNICATION Volume: 32 Issue: 1 Pages: 23-42 Published: JAN 2 2015
Times Cited: 0 *(from Web of Science Core Collection)*

2. **Immigration Judges and U.S. Asylum Policy**
By: Miller, B; Keith, LC; Holmes, JS
IMMIGRATION JUDGES AND U.S. ASYLUM POLICY Book Series: Pennsylvania Studies in Human Rights Pages: 1-238 Published: 2015
Publisher: UNIV PENNSYLVANIA PRESS, 3905 SPRUCE STREET, PHILADELPHIA, PA 19104 USA
Times Cited: 1 *(from Web of Science Core Collection)*

3. **UNITARY INNOVATIONS AND POLITICAL ACCOUNTABILITY**
By: Stiglitz, Edward H.
CORNELL LAW REVIEW Volume: 99 Issue: 5 Pages: 1133-1184 Published: JUL 2014
Times Cited: 0 *(from Web of Science Core Collection)*

4. **LITIGATION REFORM: AN INSTITUTIONAL APPROACH**
By: Burbank, Stephen B.; Farhang, Sean
UNIVERSITY OF PENNSYLVANIA LAW REVIEW Volume: 162 Issue: 7 Pages: 1543-1618 Published: JUN 2014
Times Cited: 4 *(from Web of Science Core Collection)*

5. **Does the Law Matter? Win Rates and Law Reforms**
By: Gliksberg, David
JOURNAL OF EMPIRICAL LEGAL STUDIES Volume: 11 Issue: 2 Pages: 378-407 Published: JUN 2014
Times Cited: 0 *(from Web of Science Core Collection)*

6. **Examining the Effects of Information, Attorney Capability, and Amicus Participation on U.S. Supreme Court Decision Making**
By: Szmer, John; Ginn, Martha Humphries
AMERICAN POLITICS RESEARCH Volume: 42 Issue: 3 Pages: 441-471 Published: MAY 2014
Times Cited: 0 *(from Web of Science Core Collection)*

Source: Thomson Reuters, *Web of Science* database, www.isiknowledge.com.

HELPFUL HINTS

Finding a Term on a Page

Most Internet browsers have a "hot key" combination that allows you to search for a particular word or phrase on a displayed Web page. With Internet Explorer and Firefox, for example, use CTRL-F. Take advantage of this shortcut when viewing a massive document that has small text or lots of content.

Managing Citations

Databases usually have another extremely valuable feature: the ability to electronically store citation information. In figure 3-2, to the left of the "Add to Marked List" box, is a box containing "Save to EndNote online" with an arrow for a drop-down menu. If you click on the arrow, a list of citation management options from which to choose will appear. Citation management systems store the information you need to cite your references properly even if you do not yet know what reference style you are going to use. Once you decide on the style, these systems will format your references accordingly. Furthermore, these systems usually allow you to create files to manage or sort sources into categories for use in writing your literature review.

Identifying Useful Popular Sources

As most students use the Internet on a regular basis, you may be familiar with using it to look for articles and other sources of information on topics of interest. One of the benefits of the revolution in global communications is that it places an almost limitless supply of information literally at your fingertips. Scouring the Internet also allows you to find many kinds of documents and data that a traditional library search will not turn up, or that simply are not available on many campuses. It is tempting to think that you need only to access a **search engine**, a computer program that systematically visits and searches Web pages, and type in a few search terms, or keywords. But, however powerful the facilities may be, the search process is not always simple.

Search engines such as Google or Yahoo! may be a good place to start if you are trying to see what sources are available on a topic and you are not looking for a specific reference. These search engines can be quite indiscriminate in what they return, however, and leave the user with pages of unsuitable or redundant findings. A Google Scholar search conducted on February 1, 2015, on *judicial behavior,* restricted to articles between the years 2000–2015, yielded 6,960 hits. That is, of course, way too many to read.

Search programs often order the results by the frequency of appearance of search words in the title and in the text near the top of the page, or by the regularity with which the page is visited. But these may not be the best criteria for your purposes. Use of the Internet clearly has drawbacks unless careful planning and thought have preceded the search. As with many of life's activities, the more time you spend searching for literature review materials, the easier it will become. Nevertheless, following a few practical guidelines will expedite the process.

- When first visiting a site, particularly one with search features, click the "Help" button, which usually provides specific instructions for how to search that site.

- If possible, pyramid your search by going first to a political science page and, from there, looking for more specific sites.
- If you have a clear topic in mind, start with a specific Internet site, such as one sponsored by a research organization or university. Doing so will reduce the number of false hits.
- Open a simple word-processing program such as Notepad or WordPad. Highlight and copy selected text from a Web page to facilitate collecting information. Be sure to document the source of this material properly. This technique is especially helpful for copying complicated, long Internet addresses (URLs).
- On a complicated page with lots of text and images, use your browser's "Find" option to locate the word or phrase of interest.
- Take advantage of advanced search options. If possible, limit your search to specified periods, to certain types of articles, to particular authors or subjects, and to data formats.
- Check this book's Web site (http://psrm.cqpress.com/) for links to specific topics.

Most search engines and databases enable you to narrow a search to meet your specific needs. Usually, you want to see only the documents that contain all the words—or even specific phrases, such as *international terrorism*—on a list. Advanced search features allow you to use connectors and modifiers to specify exactly what words or phrases should be included in the document and which ones should be excluded. If you enter the desired words without adding modifiers, in all likelihood the search engine will look for pages that contain any of the listed words but not necessarily all of them.

Although the Internet allows for a wide search of material, not all information found on the Internet is reliable. Virtually anyone or any group, no matter what its credentials are, can create a Web site. The only way to know for sure that the information you are looking at is dependable is to be familiar with the site's sponsor. In general, sites presented by individuals, even those with impressive-looking titles and qualifications, may not have the credibility or scholarly standing that your literature review requires. In contrast, you can usually have confidence in sources cited in professional publications or by established authors or reputable organizations. Note, too, that even sources in the form of opinion can be dependable. Many associations that hold strong political or ideological positions nevertheless offer useful information that is worth citing. If in doubt about the reliability of a source, check with your instructor or adviser. He or she should be able to help you assess whether or not accessed information is usable.

Internet sources must be cited properly, partly because so much variation exists in the quality of these sources but also, and even more important, because academic standards dictate that proper citations be provided for any work consulted. In this way authors are fully credited for their data and ideas, and readers can check the accuracy of the information and the quality of a literature review.

At a minimum, the citation should include the author or creator of the page and the title of the article, as well as the complete Internet address at which the article was found. If the information you retrieve from a Web site is likely to have changed since you accessed it, as in the case of a crowdsourced encyclopedia article or a page that continuously posts up-to-date data, then you would add the date you accessed the site, perhaps in parentheses after the URL. Following is a generic format for citing a Web page in a bibliography:

Author [last name–first name or full organization name]. (Date of publication, if available). Web Page Title. Full Web address.

For example,

Stroupe, Kenneth S., Jr., & Larry J. Sabato. (2004). Politics: The Missing Link of Responsible Civic Education (CIRCLE working paper 18). http://www.civic youth.org/PopUps/WorkingPapers/WP18Stroupe.pdf

indicates that your information is from a report by Kenneth S. Stroupe Jr. and Larry J. Sabato and is available on a Web page administered by the Center for Information and Research on Civic Learning and Engagement (CIRCLE) that you accessed at http://www.civicyouth.org/PopUps/WorkingPapers/WP18Stroupe.pdf.

Citation style will depend on the standards set by your institution or instructor, but include at least enough detail to let a reader retrieve the page and verify information. We have included a number of guides for citing references and conducting and writing literature reviews in the "Suggested Readings" section at the end of this chapter.

Reading the Literature

Once you have identified references for possible inclusion in a literature review, the next step is to figure out how the references fit together in a way that (1) explains the base of knowledge, or what we know about a topic from previous work, with respect to the research question, and (2) establishes how the current project is going to build on that knowledge. The best way to understand the base of knowledge is to read the work that answers the central research questions and understand how each contributes to a comprehensive understanding of the important research questions. To read an entire literature would take far too much time, so it is wise to rely on shortcuts whenever available.

First, following the suggestions in the preceding section, take care in selecting references. Once references are identified and collected, you can rely on the abstract on the first page of most articles and the preface at the beginning of most books to serve as a short description of the whole work and the conclusions contained therein. A good abstract will include a great deal of important information about the contents of an article, including the research question, the theory and hypotheses, the data and methods used to test the hypotheses, and the results and conclusions. Most article abstracts are only two hundred to three hundred words long, so they offer an easy way to assess quickly whether an article is worth reading further. A good preface will include the same kind of information, but a book's length makes this summary much more cursory or general. A preface will also include more attention to organization of the chapters. Reading book reviews in scholarly journals is another way to learn quickly the value of a book to a given project. For most books, you can find a review that will relay the book's theoretical importance or help you understand how it fits in the context of the existing literature and what it adds to the base of knowledge—in addition to assessing the quality of the research.

Use of abstracts, prefaces, and book reviews will help narrow a list of references. This smaller list can then be culled for those references that are essential to motivating the current research project and those that add depth, range, or a unique perspective to the literature review. In addition, the first few pages of political science articles contain most of the description of the key components of the research project—the research question, theory and hypotheses, and data and methods—and include a literature review. The conclusion or discussion of findings will summarize the results and explain how they add to the base of knowledge. Students with limited time for reading articles should read the first few pages and the conclusion and then, if more information is needed, proceed to the rest of the article. Finally, although many political science articles include complex methods and tables, the text describing the results usually includes a more jargon-free description of the results that does not require an advanced understanding of statistics. The same time-saving tips can be applied to books by concentrating on a book's introduction and conclusion as well as selected relevant chapters, which you can identify in the table of contents.

Nonscholarly references like magazine or newspaper articles, or Web site content, generally are much shorter than references from the scholarly literature and require fewer shortcuts. These sources can typically be read quickly, and in most cases do not provide an abstract.

Writing a Literature Review

After you have identified the relevant literature and started reading the literature, it is time to begin crafting the literature review. In this section, we explain how you can integrate a collection of related materials into an effective literature review. Essential to this process is limiting the discussion of materials to the most relevant previous work and focusing the literature review on concepts and ideas rather than around individual books, articles, or authors. This is important because organizing the literature review in this way will make it easier to establish the base of knowledge and demonstrate how the current research project can extend or add to that knowledge—with a new perspective, new data, or a different method—by resolving conflicting results in the literature or by replicating, and thereby validating, previous research. When thinking about a literature review as motivating a research project in one of these ways, you will see that the literature review is an integral part of a research project and requires a great deal of attention to establish the direction of the project.

The key to organizing and writing an effective literature review is to focus on concepts, ideas, and methods shared across the literature. Many students are used to writing about multiple references with a focus on the individual references, discussing each collected reference in turn. For example, imagine that you have collected ten articles for a literature review. You might decide that the easiest way to organize a review that incorporates all ten articles would be to take the first article, perhaps selected because it was the most relevant, and summarize the most important parts of the article: the research question, theory, hypotheses, data, methods, and results. After summarizing the first article, you then move on to the second article and write a similar summary in the next paragraph, then the third, and so on, until all ten articles have been summarized. We call this approach to a literature review the "boxcar method" because such a review links the independent discussions of each article much like a series of boxcars on a train.

Although this may be the easiest method for including multiple references in a literature review, it is ineffective. It does not explain how the ten articles fit together to establish the base of knowledge to which the current project will add; nor does it establish how the current project will add to that knowledge. By tacking together independent discussions of articles, you will find it difficult to discuss common themes across references, conflicting results or conclusions, or questions left unanswered in the literature.

A more effective way to write a literature review is to focus on the concepts, ideas, and methods in the relevant literature. Think of a literature review as an essay about what has been written on your topic. You are most likely already familiar with doing this if you have written a research paper that did not include your own data analysis. In this case, however, your essay is going to focus on themes and concepts related to your research and your analysis of data. For example, imagine that you have the same ten articles from the previous example, but instead of discussing each independently, you begin by identifying the common themes across all ten articles. The first step might be to group the articles according to their research questions.

It is likely that all ten articles address a similar broad topic but do not share exactly the same research questions. You can begin to establish the base of knowledge by identifying, for example, three common research questions among the ten articles (four articles answering question one, three articles answering question two, and three articles answering question three). These three research questions represent the three areas of study the previous literature has undertaken in building our understanding of the broader topical area. Beginning the literature review with a discussion of these three research questions, and citing the articles that use each, will be an effective start to defining the base of knowledge.

Next, you might regroup the articles based on the data or research designs used. Perhaps three of the articles used experiments, and seven of the articles used case studies. Researchers commonly discuss in their literature reviews the different research designs used in the literature because, as explained in chapter 6, different approaches have advantages and disadvantages and will be better or worse for making certain kinds of conclusions.

In addition to differences in research design, each of the three experiments and seven case studies also likely used different data. Some of the case studies may have relied on personal interviews; others may have used participant-observation methods. Likewise, some of the experiments may have collected data from college students, and others may have collected data from the general population. Differences in the method of collecting the data or the populations from which the data were collected might lead to different conclusions.

As a final example, you might sort the ten references by the results or conclusions. It is unlikely that all ten articles came to the same conclusion. In fact, the results of at least one of the ten articles likely contradict the results of the others. Identifying commonalities and contradictions in the literature review allows a researcher to identify ideas that have been established through replication as accepted widely in the literature and areas of disagreement that are ripe for further clarification and explanation. Conflicting results can provide a wonderful motivating factor for new research and establish for the reader the importance and relevance of the current research project.

Compared to the boxcar method, the latter example describes a much more sophisticated literature review because it integrates previous research along conceptual and methodological lines and provides a more effective organization for the researcher to explain the base of knowledge and how the current project fits into that literature. As we noted earlier, the boxcar method may be attractive because it seems easier, but the integrated literature review will better inform the current research project and the reader—and, practically speaking, will earn a better grade for students.

A literature review is not all that different from a conventional research paper in which you write an essay about what is known about a topic. In both cases, the discussion needs to be organized around key themes, and it is your task as the reviewer to choose the important themes on which to focus. A literature review for an empirical research paper tends to focus more on methodological aspects of previous studies in addition to the substantive content of previous research.

Anatomy of a Literature Review

To demonstrate further how you might write a highly effective literature review, we include in figure 3-4 a literature review from an article discussed in chapter 1: "Does Attack Advertising Demobilize the Electorate?" by Stephen Ansolabehere, Shanto Iyengar, Adam Simon, and Nicholas Valentino. In this section, we dissect this literature review to highlight the value of integrating references by focusing on concepts and ideas rather than individual articles or books. This literature review begins with the first paragraph in the article and continues to page 2. As we will see, the authors do an excellent job of explaining previous work on the effect of campaign advertising on voters, explaining the received wisdom from this work, identifying the shortcomings of previous work, and explaining how this article will correct those shortcomings.

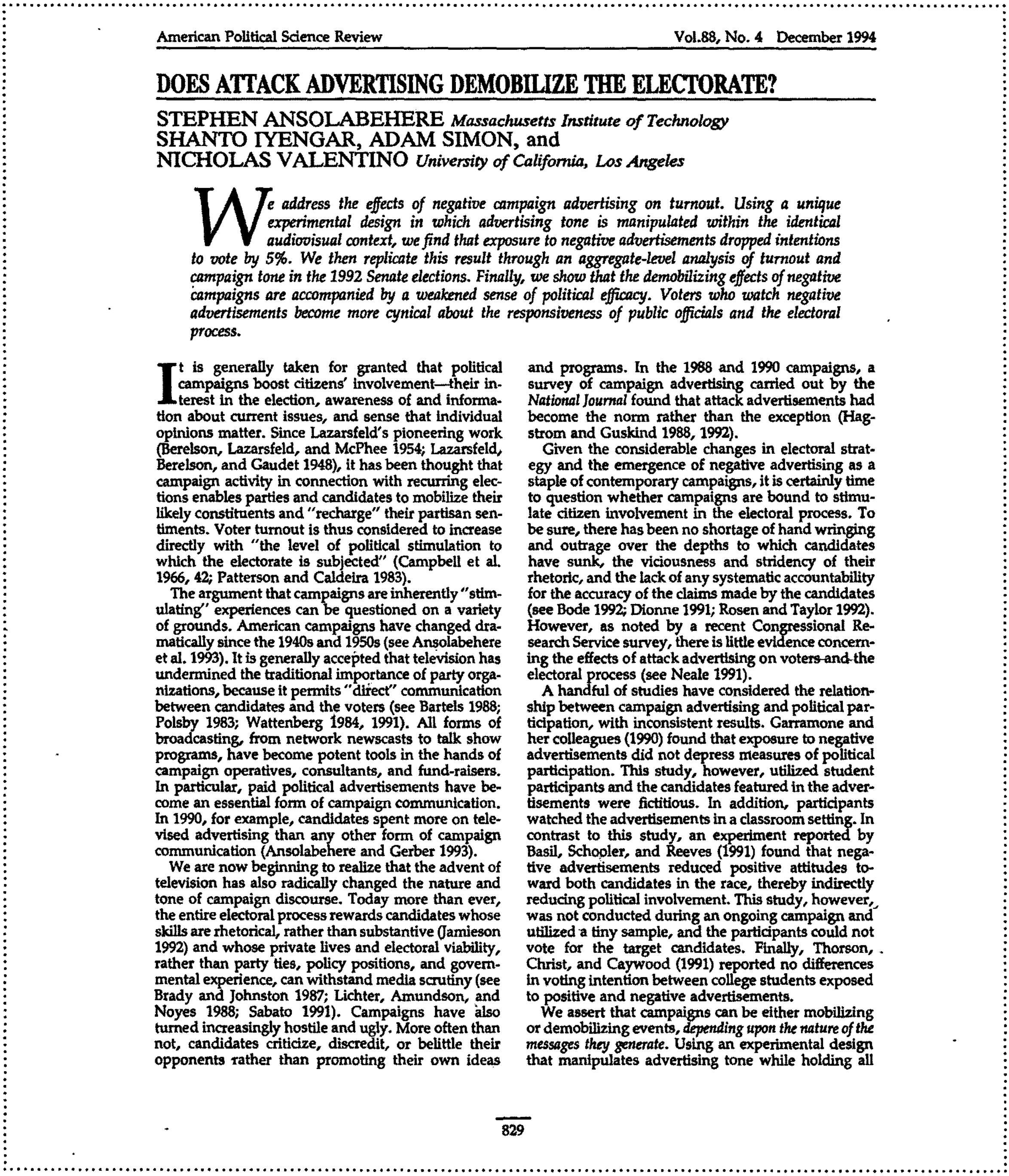
Note first that this is a scholarly article from a highly respected political science journal. The article is written following the style and citation guidelines for the *American Political Science Review* (*APSR*). *APSR* and many other journals use parenthetical notation to identify for the reader, at a glance, the names of the cited authors, the year of the cited publication, and a page number if relevant. The interested reader will find that the names and dates match a full citation in the works cited at the end of the literature review. Other journals may use a different citation style, such as endnotes or footnotes, but in all cases the author must provide citations acknowledging others' work and a full citation within the article. You should do the same, or your literature review will fail to give credit where credit is due and leave you open to charges of plagiarism.

In the first paragraph, the authors begin by identifying the conventional wisdom that "it is generally taken for granted that political campaigns boost citizens' involvement—their interest in the election, awareness of and information about current issues, and sense that individual opinions matter."[4] This sentence succinctly captures the essence of the received wisdom about the relationship between

(continued on p. 100)

[4] Stephen Ansolabehere, Shanto Iyengar, Adam Simon, and Nicholas Valentino, "Does Attack Advertising Demobilize the Electorate?" *American Political Science Review* 88, no. 4 (1994): 829. Available at http://weber.ucsd.edu/~tkousser/Ansolabehere.pdf

FIGURE 3-4 A Well-Constructed Literature Review



American Political Science Review Vol.88, No. 4 December 1994

DOES ATTACK ADVERTISING DEMOBILIZE THE ELECTORATE?

STEPHEN ANSOLABEHERE *Massachusetts Institute of Technology*
SHANTO IYENGAR, ADAM SIMON, and
NICHOLAS VALENTINO *University of California, Los Angeles*

We address the effects of negative campaign advertising on turnout. Using a unique experimental design in which advertising tone is manipulated within the identical audiovisual context, we find that exposure to negative advertisements dropped intentions to vote by 5%. We then replicate this result through an aggregate-level analysis of turnout and campaign tone in the 1992 Senate elections. Finally, we show that the demobilizing effects of negative campaigns are accompanied by a weakened sense of political efficacy. Voters who watch negative advertisements become more cynical about the responsiveness of public officials and the electoral process.

It is generally taken for granted that political campaigns boost citizens' involvement—their interest in the election, awareness of and information about current issues, and sense that individual opinions matter. Since Lazarsfeld's pioneering work (Berelson, Lazarsfeld, and McPhee 1954; Lazarsfeld, Berelson, and Gaudet 1948), it has been thought that campaign activity in connection with recurring elections enables parties and candidates to mobilize their likely constituents and "recharge" their partisan sentiments. Voter turnout is thus considered to increase directly with "the level of political stimulation to which the electorate is subjected" (Campbell et al. 1966, 42; Patterson and Caldeira 1983).

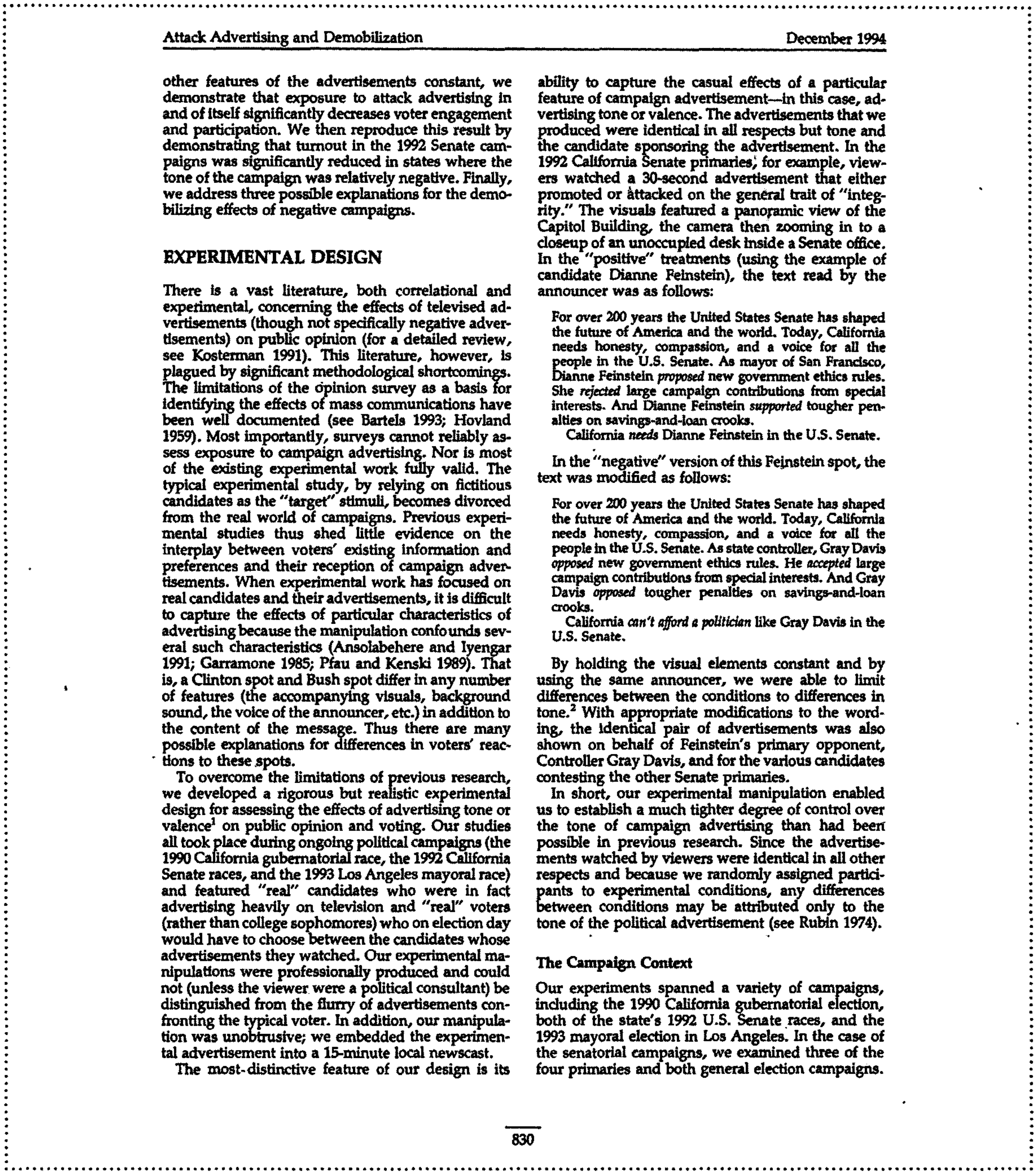
The argument that campaigns are inherently "stimulating" experiences can be questioned on a variety of grounds. American campaigns have changed dramatically since the 1940s and 1950s (see Ansolabehere et al. 1993). It is generally accepted that television has undermined the traditional importance of party organizations, because it permits "direct" communication between candidates and the voters (see Bartels 1988; Polsby 1983; Wattenberg 1984, 1991). All forms of broadcasting, from network newscasts to talk show programs, have become potent tools in the hands of campaign operatives, consultants, and fund-raisers. In particular, paid political advertisements have become an essential form of campaign communication. In 1990, for example, candidates spent more on televised advertising than any other form of campaign communication (Ansolabehere and Gerber 1993).

We are now beginning to realize that the advent of television has also radically changed the nature and tone of campaign discourse. Today more than ever, the entire electoral process rewards candidates whose skills are rhetorical, rather than substantive (Jamieson 1992) and whose private lives and electoral viability, rather than party ties, policy positions, and governmental experience, can withstand media scrutiny (see Brady and Johnston 1987; Lichter, Amundson, and Noyes 1988; Sabato 1991). Campaigns have also turned increasingly hostile and ugly. More often than not, candidates criticize, discredit, or belittle their opponents rather than promoting their own ideas and programs. In the 1988 and 1990 campaigns, a survey of campaign advertising carried out by the *National Journal* found that attack advertisements had become the norm rather than the exception (Hagstrom and Guskind 1988, 1992).

Given the considerable changes in electoral strategy and the emergence of negative advertising as a staple of contemporary campaigns, it is certainly time to question whether campaigns are bound to stimulate citizen involvement in the electoral process. To be sure, there has been no shortage of hand wringing and outrage over the depths to which candidates have sunk, the viciousness and stridency of their rhetoric, and the lack of any systematic accountability for the accuracy of the claims made by the candidates (see Bode 1992; Dionne 1991; Rosen and Taylor 1992). However, as noted by a recent Congressional Research Service survey, there is little evidence concerning the effects of attack advertising on voters and the electoral process (see Neale 1991).

A handful of studies have considered the relationship between campaign advertising and political participation, with inconsistent results. Garramone and her colleagues (1990) found that exposure to negative advertisements did not depress measures of political participation. This study, however, utilized student participants and the candidates featured in the advertisements were fictitious. In addition, participants watched the advertisements in a classroom setting. In contrast to this study, an experiment reported by Basil, Schooler, and Reeves (1991) found that negative advertisements reduced positive attitudes toward both candidates in the race, thereby indirectly reducing political involvement. This study, however, was not conducted during an ongoing campaign and utilized a tiny sample, and the participants could not vote for the target candidates. Finally, Thorson, Christ, and Caywood (1991) reported no differences in voting intention between college students exposed to positive and negative advertisements.

We assert that campaigns can be either mobilizing or demobilizing events, *depending upon the nature of the messages they generate*. Using an experimental design that manipulates advertising tone while holding all



other features of the advertisements constant, we demonstrate that exposure to attack advertising in and of itself significantly decreases voter engagement and participation. We then reproduce this result by demonstrating that turnout in the 1992 Senate campaigns was significantly reduced in states where the tone of the campaign was relatively negative. Finally, we address three possible explanations for the demobilizing effects of negative campaigns.

EXPERIMENTAL DESIGN

There is a vast literature, both correlational and experimental, concerning the effects of televised advertisements (though not specifically negative advertisements) on public opinion (for a detailed review, see Kosterman 1991). This literature, however, is plagued by significant methodological shortcomings. The limitations of the opinion survey as a basis for identifying the effects of mass communications have been well documented (see Bartels 1993; Hovland 1959). Most importantly, surveys cannot reliably assess exposure to campaign advertising. Nor is most of the existing experimental work fully valid. The typical experimental study, by relying on fictitious candidates as the "target" stimuli, becomes divorced from the real world of campaigns. Previous experimental studies thus shed little evidence on the interplay between voters' existing information and preferences and their reception of campaign advertisements. When experimental work has focused on real candidates and their advertisements, it is difficult to capture the effects of particular characteristics of advertising because the manipulation confounds several such characteristics (Ansolabehere and Iyengar 1991; Garramone 1985; Pfau and Kenski 1989). That is, a Clinton spot and Bush spot differ in any number of features (the accompanying visuals, background sound, the voice of the announcer, etc.) in addition to the content of the message. Thus there are many possible explanations for differences in voters' reactions to these spots.

To overcome the limitations of previous research, we developed a rigorous but realistic experimental design for assessing the effects of advertising tone or valence[1] on public opinion and voting. Our studies all took place during ongoing political campaigns (the 1990 California gubernatorial race, the 1992 California Senate races, and the 1993 Los Angeles mayoral race) and featured "real" candidates who were in fact advertising heavily on television and "real" voters (rather than college sophomores) who on election day would have to choose between the candidates whose advertisements they watched. Our experimental manipulations were professionally produced and could not (unless the viewer were a political consultant) be distinguished from the flurry of advertisements confronting the typical voter. In addition, our manipulation was unobtrusive; we embedded the experimental advertisement into a 15-minute local newscast.

The most-distinctive feature of our design is its ability to capture the casual effects of a particular feature of campaign advertisement—in this case, advertising tone or valence. The advertisements that we produced were identical in all respects but tone and the candidate sponsoring the advertisement. In the 1992 California Senate primaries, for example, viewers watched a 30-second advertisement that either promoted or attacked on the general trait of "integrity." The visuals featured a panoramic view of the Capitol Building, the camera then zooming in to a closeup of an unoccupied desk inside a Senate office. In the "positive" treatments (using the example of candidate Dianne Feinstein), the text read by the announcer was as follows:

> For over 200 years the United States Senate has shaped the future of America and the world. Today, California needs honesty, compassion, and a voice for all the people in the U.S. Senate. As mayor of San Francisco, Dianne Feinstein *proposed* new government ethics rules. She *rejected* large campaign contributions from special interests. And Dianne Feinstein *supported* tougher penalties on savings-and-loan crooks.
>
> California *needs* Dianne Feinstein in the U.S. Senate.

In the "negative" version of this Feinstein spot, the text was modified as follows:

> For over 200 years the United States Senate has shaped the future of America and the world. Today, California needs honesty, compassion, and a voice for all the people in the U.S. Senate. As state controller, Gray Davis *opposed* new government ethics rules. He *accepted* large campaign contributions from special interests. And Gray Davis *opposed* tougher penalties on savings-and-loan crooks.
>
> California *can't afford a politician* like Gray Davis in the U.S. Senate.

By holding the visual elements constant and by using the same announcer, we were able to limit differences between the conditions to differences in tone.[2] With appropriate modifications to the wording, the identical pair of advertisements was also shown on behalf of Feinstein's primary opponent, Controller Gray Davis, and for the various candidates contesting the other Senate primaries.

In short, our experimental manipulation enabled us to establish a much tighter degree of control over the tone of campaign advertising than had been possible in previous research. Since the advertisements watched by viewers were identical in all other respects and because we randomly assigned participants to experimental conditions, any differences between conditions may be attributed only to the tone of the political advertisement (see Rubin 1974).

The Campaign Context

Our experiments spanned a variety of campaigns, including the 1990 California gubernatorial election, both of the state's 1992 U.S. Senate races, and the 1993 mayoral election in Los Angeles. In the case of the senatorial campaigns, we examined three of the four primaries and both general election campaigns.

Source: Stephen Ansolabehere, Shanto Iyengar, Adam Simon, and Nicholas Valentino, "Does Attack Advertising Demobilize the Electorate?" *American Political Science Review* 88, no. 4 (Dec. 1994): 829–30. Copyright © 1994 American Political Science Association. Reprinted with the permission of Cambridge University Press.

campaigns and voters and is followed by citations of those responsible for laying the early groundwork in developing this understanding. The second and third sentences extend the discussion of the conventional wisdom and cite two more recent studies that tested these ideas and found similar results.

The second paragraph explains that the authors question this conventional wisdom and cites various changes to the nature of campaigns since the 1940s—primarily the role of television. As in the first paragraph, after introducing a new idea in the literature review, the authors include parenthetical notes citing the work responsible for the idea. In this section, the authors cite four references for the role television has played and one reference that documents the increasing importance of paid political advertising to campaign operatives.

The third paragraph discusses similar themes and cites work that examines the value of rhetorical skill and the ability to withstand media scrutiny during an election. Finally, the third paragraph explains that campaigns have become "increasingly hostile and ugly" and cites two references to establish the point. As you can see, the first three paragraphs of this literature review are organized around concepts and ideas that are essential to understanding the base of knowledge about the relationship between campaign advertising and voters.

An important aspect of the fourth and fifth paragraphs is that they transition from establishing that the nature of campaigns has changed since early work on the topic to establishing that some work has attempted to measure this new relationship. The authors cite "Neale 1991" when claiming that "there is little evidence concerning the effects of attack advertising on voters and the electoral process."[5] They also cite three studies that examined the same research question as Ansolabehere et al.: "Garramone and her colleagues (1990)"; "Basil, Schooler, and Reeves (1991)"; and "Thorson, Christ, and Caywood (1991)." According to the authors, the previous work was inconclusive because it found conflicting results. Garramone et al. found that negative advertising did not depress turnout; Basil, Schooler, and Reeves found that negative advertisements indirectly reduced political participation; and Thorson, Christ, and Caywood reported that negative advertisements had no effect on the intention of voting. With each citation, the authors also identify some of the problems in each research design that might lead to suspect results. Given these conflicting results, the authors propose in the sixth paragraph that they will attempt to provide clarity by improving upon previous work by correcting research design flaws.

The first new paragraph on the second page, under the "Experimental Design" heading, provides further detail about the flaws of previous work using two different approaches: survey research and experimental research. The authors first point

[5] Ibid.

the interested reader to another reference that has documented the literature on television advertising and public opinion, "Kosterman 1991." They then turn their attention to survey research and identify the main drawback of this approach: a lack of measurement of direct exposure to advertising, as documented by two cited references. Next, the authors discuss the flaws of previous experimental work, primarily issues of external validity, and point to three cited references. The following paragraph begins the description of this article's research design.

With this example, you can see that there is a logical order to the literature review: establish conventional wisdom, establish that the nature of politics has changed—while the conventional understanding has not, and identify flaws in previous research that can be corrected. Discussing the literature in this manner makes a convincing case to the reader that this research project will be an important addition to the literature because it will improve our understanding of a topic that until now has been misunderstood.

Also, by organizing the literature review in this way, the authors have found a clear motivation for designing their research project as they have. Throughout the literature review, the authors integrated twenty-nine references by focusing on the concepts, ideas, and methods that were shared across the literature.

Finally, the authors established that this is an important area of study (as others have an interest in writing in this area), and that our understanding is not complete (as there is disagreement through conflicting results and conclusions).

Although different literature reviews will vary in the organizational style they use, we recommend that students working on their own literature reviews try to follow this topical style of integrating references; it will make even a brief discussion, like the two pages in the Ansolabehere et al. article, very powerful.

Conclusion

No matter what the original purpose of your literature review may have been, it should be thorough. In your research report, you should discuss the sources that provide explanations for the phenomenon you are studying and that support the plausibility of your hypotheses. You should also discuss how your research relates to other research and use the existing literature to document the significance of your research. You can look to the example in the previous section or to an example of a literature review contained in the research report in chapter 15. Another way to learn about the process is to read a few articles in any of the main political science journals that we listed earlier in this chapter and take some time to study the literature reviews carefully, looking for effective styles that would suit your own project.

TERMS INTRODUCED

Electronic databases. A collection of information (of any type) stored on an electromagnetic medium that can be accessed and examined by certain computer programs.

Literature review. A systematic examination and interpretation of the literature for the purpose of informing further work on a topic.

Relationship. The association, dependence, or covariance of the values of one variable with the values of another.

Search engine. A computer program that visits Web pages on the Internet and looks for those containing particular directories or words.

Search term. A word or phrase entered into a computer program (a search engine) that looks through Web pages on the Internet for those that contain the word or phrase.

SUGGESTED READINGS

Fink, Arlene. *Conducting Research Literature Reviews: From the Internet to Paper.* 4th ed. Thousand Oaks, Calif.: Sage, 2014.

Galvan, Jose L. *Writing Literature Reviews: A Guide for Students of the Social and Behavioral Sciences.* 5th ed. Glendale, Calif.: Pyrczak, 2013.

Williams, Kristen. *Research and Writing Guide for Political Science.* New York: Oxford University Press, 2014.

CHAPTER

4

The Building Blocks of Social Scientific Research:

Hypotheses, Concepts, and Variables

CHAPTER OBJECTIVES

4.1 Identify the types of variables involved in an explanation for a phenomenon.

4.2 Explain the characteristics of good hypotheses.

4.3 Discuss the role of defining concepts in research studies.

IN CHAPTERS 1 AND 2, WE DISCUSSED what it means to acquire scientific knowledge and presented examples of political science research intended to produce this type of knowledge. In chapter 3, we discussed how to search for a topic and begin to pose an appropriate research question. In this chapter, we focus on taking the next steps beyond specifying the research question. These steps require us to (1) propose a suitable explanation for the phenomena under study, (2) formulate testable hypotheses, and (3) define the concepts identified in the hypotheses. Although we discuss these steps as if they occur in sequence, the actual order may vary. All the steps must be taken eventually, however, before a research project can be completed successfully. The sooner the issues and decisions involved in each of the steps are addressed, the sooner the other portions of the research project can be completed.

Proposing Explanations

Once a researcher has developed a suitable research question or topic, the next step is to propose an explanation for the phenomenon the researcher is interested in understanding. Proposing an explanation involves identifying other phenomena that we think will help us account for the object of our research and then specifying how and why these two (or more) phenomena are related. Or, alternatively, we may identify a political phenomenon and want to know whether or not it has any impact on other political phenomena. Developing an explanation involves thinking about relationships between concepts. Your literature review should give you plenty of ideas about relationships between concepts.

In the examples referred to in chapter 1, the researchers proposed explanations for the political phenomena they were studying. Kenworthy and Pontusson investigated whether increases in inequality of market incomes lead to increases in government spending for redistributive programs.[1] Hajinal and Horowitz investigated whether minorities fared better when Republicans were in control of the federal government or when Democrats were in control.[2] Minkler and Sweeney investigated whether developing countries respect security and subsistence rights simultaneously.[3] Nicholson and Hansford wanted to know whether partisanship influenced public support for Supreme Court decisions.[4] Dowling and Wichowsky wanted to know if revealing the identity of sponsors of negative campaign ads changed the impact of those ads on the public.[5] And Fox and Lawless wanted to know what factors account for the gender gap in political ambition.[6]

To help clarify relationships between phenomena, political scientists refer to phenomena as variables and identify several types of variables. A phenomenon that we think will help us explain political characteristics or behavior is called an **independent variable**. Independent variables are thought to influence, affect, or cause some other phenomenon. A **dependent variable** is thought to be caused, to depend

[1] Lane Kenworthy and Jonas Pontusson, "Rising Inequality and the Politics of Redistribution in Affluent Countries," *Perspectives on Politics* 3, no. 3 (2005): 449–71. Available at http://www.u.arizona.edu/~lkenwor/pop2005.pdf

[2] Zoltan L. Hajinal and Jeremy D. Horowitz, "Racial Winners and Losers in American Party Politics," *Perspectives on Politics* 12, no.1 (2014): 110–18.

[3] Lanse Minkler and Shawna Sweeny, "On the Indivisibility and Interdependence of Basic Rights in Developing Countries," *Human Rights Quarterly* 33 (2011): 351–96.

[4] Stephen P. Nicholson and Thomas G. Hansford, "Partisans in Robes: Party Cues and Public Acceptance of Supreme Court Decisions," *American Journal of Political Science* 58, no. 3 (2014): 620–36.

[5] Conor M. Dowling and Amber Wichowsky, "Does It Matter Who's Behind the Curtain? Anonymity in Political Advertising and the Effects of Campaign Finance Disclosure," *American Politics Research* 41, no. 6 (2013): 965–96.

[6] Richard L. Fox and Jennifer L. Lawless, "Uncovering the Origins of the Gender Gap in Political Ambition," *American Political Science Review*, 108, no. 3 (2014): 499–519.

upon, or to be a function of an independent variable. Thus, if a researcher has hypothesized that acquiring more formal education will lead to increased income later on (in other words, that income may be explained by education), then years of formal education would be the independent variable, and income would be the dependent variable.

As the word *variable* connotes, we expect the value of the concepts we identify as variables to vary or change. A concept that does not change in value is called a **constant** and cannot be used to investigate a relationship. Unfortunately, sometimes a concept is expected to vary and thus be suitable for inclusion in a research project, only for a researcher to discover later that the concept does not vary in the context in which it is being used. For example, a student working on a survey to be distributed to her classmates wanted to see if students having served in the military or having a family member in the military had different attitudes toward the war in Iraq than did students without military service connections. She discovered that none of the students had any military service connections: having military service connections was a constant. As a result she had to think of other factors that might account for differences in student attitudes toward the war in Iraq.

Proposed explanations for political phenomena are often more complicated than the simple identification of one independent variable that is thought to explain variation in a dependent variable. More than one phenomenon is usually needed to account adequately for most political behavior. For example, suppose a researcher proposes the following relationship between state efforts to regulate pollution and the severity of potential harm from pollution: the higher the threat of pollution (independent variable), the greater the effort to regulate pollution (dependent variable). The insightful researcher would realize the possibility that another phenomenon, such as the wealth of a state, might also affect a state's regulatory effort. As another example, remember from chapter 1 that Lane Kenworthy and Jonas Pontusson thought that larger changes in market inequality would cause larger changes in redistribution but that changes in redistribution would also be affected by turnout rates in national elections.[7] In later chapters we will discuss how one measures the impact of independent variables, individually and in combination, on a dependent variable. Sometimes, in addition to proposing that independent variables are related to the dependent variable, researchers propose relationships between the independent variables. In particular, we might want to determine which independent variables occur before other independent variables and indicate which ones have a more direct, as opposed to indirect, effect on the phenomenon we are trying to explain (the dependent variable). A variable that occurs prior to all other variables and that may affect other independent variables is called an **antecedent variable.** A variable that occurs closer in time to the dependent variable and is itself

[7] Kenworthy and Pontusson, "Rising Inequality and the Politics of Redistribution."

affected by other independent variables is called an **intervening variable.** Consider these examples.

Suppose a researcher hypothesizes that a person who favored national health insurance was more likely to have voted for Barack Obama in 2008 than was a person who did not favor such extensive coverage. In this case, the attitude toward national health insurance would be the independent variable and the presidential vote the dependent variable. The researcher might wonder what causes the attitude toward national health insurance and might propose that those people who have inadequate medical insurance are more apt to favor national health insurance. This new variable (adequacy of a person's present medical insurance) would then be an antecedent variable, since it comes before and affects (we think) the independent variable. Thinking about antecedent variables pushes our explanatory scheme further back in time and, we hope, will lead to a more complete understanding of a particular phenomenon (in this case, presidential voting). Notice how the independent variable in the original hypothesis (attitude toward national health insurance) becomes the dependent variable in the hypothesis involving the antecedent variable (adequacy of health insurance). Also notice that in this example, adequacy of health insurance is thought to exert an indirect effect on the dependent variable (presidential voting) via its impact on attitudes toward national health insurance.

Now consider a second example. Suppose a researcher hypothesizes that a voter's years of formal education affect her or his propensity to vote. In this case, education would be the independent variable and voter turnout the dependent variable. If the researcher then begins to consider what about education has this effect, he or she has begun to identify the intervening variables between education and turnout. For example, the researcher might hypothesize that formal education creates or causes a sense of civic duty, which in turn encourages voter turnout, or that formal education causes an ability to understand the different issue positions of the candidates, which in turn causes voter turnout. Intervening variables come between an independent variable and a dependent variable and help explain the process by which one influences the other.

Explanatory schemes that involve numerous independent, alternative, antecedent, and intervening variables can become quite complex. An **arrow diagram** is a handy device for presenting and keeping track of such complicated explanations. The arrow diagram specifies the phenomena of interest; indicates which variables are independent, alternative, antecedent, intervening, and dependent; and shows which variables are thought to affect which other ones. In figure 4-1 we present arrow diagrams for the two voting examples we just considered.

In both diagrams, the dependent variable is placed at the end of the time line, with the independent, alternative, intervening, and antecedent variables placed in their appropriate locations to indicate which ones come earlier and which come later.

Arrows indicate that one variable is thought to explain or be related to another; the direction of the arrow indicates which variable is independent and which is dependent in that proposed relationship.

Figure 4-2 shows two examples of arrow diagrams that have been proposed and tested by political scientists. Both diagrams are thought to explain presidential voting behavior. In the first diagram, the ultimate dependent variable, Vote, is thought to be explained by Candidate Evaluations and Party Identification. The Candidate Evaluations variable, in turn, is explained by the Issue Losses, Party Identification, and Perceived Candidate Personalities variables. These, in turn, are explained by other concepts in the diagram. The variables at the top of the diagram tend to be antecedent variables (the subscript *t*–1 denotes that these variables precede variables with subscript *t*, where *t* indicates time); the ones in the center tend to be intervening variables. Nine independent variables of one sort or another figure in the explanation of the vote.

The second diagram also has Vote as the ultimate dependent variable, which is explained directly by only one independent variable, Comparative Candidate Evaluations. The latter variable, in turn, is dependent upon six independent variables: Personal Qualities Evaluations, Comparative Policy Distances, Current Party Attachment, Region, Religion, and Partisan Voting History. In this diagram, sixteen variables figure, either indirectly or directly, in the explanation of the Vote variable, with the antecedent variables located around the perimeter of the diagram and the intervening variables closer to the center. Both of these diagrams clearly represent complicated and extensive attempts to explain a dependent variable.

Note that arrow diagrams show hypothesized causal relationships. A one-headed arrow connecting two variables is a shorthand way of expressing the proposition "*X* directly causes *Y*." If arrows do not directly link two variables, the variables may be associated or correlated, but the relationship is indirect, not causal. As we discuss in greater depth in chapter 6, when we assert *X* causes *Y*, we are in effect making three claims. One is that *X* and *Y* covary—a change in one variable is associated with a change in the other. Also, we are claiming that a change in the independent variable (*X*) *precedes* the change in the dependent variable (*Y*). Finally, we are stating that the covariation between *X* and *Y* is not simply a coincidence or spurious—that is, due to change in some other variable—but is direct.

We have discussed the first two steps in the research process—asking a question and then proposing an explanation by suggesting how concepts or variables are related to one another—as occurring in this order, but quite often this is not the case. Researchers might start out with a theory and make deductions based on it. In other words, they start with an explanation and look for an appropriate research question that the theory might answer. Theory is an important aspect of explanation, for in order to be able to argue effectively that something causes something else, we need to be able to supply a reason or, to use words from the natural sciences, to identify the *mechanism* behind the relationship. This is the role of theory. For example, the theory of the median voter supplies a reason for changes in government policies.

Formulating Hypotheses

A **hypothesis** is an explicit statement that indicates how a researcher thinks phenomena of interest (variables) are related. It proposes a relationship that subsequently will be tested with empirical observations of the variables. A hypothesis is a guess (but of an educated nature) that indicates how an independent variable is thought to affect, influence, or alter a dependent variable. Since hypotheses are proposed relationships, they may turn out to be incorrect and not supported by the empirical evidence.

FIGURE 4-1 **Arrow Diagram of Adequacy of Medical Insurance and Voter Turnout Examples**

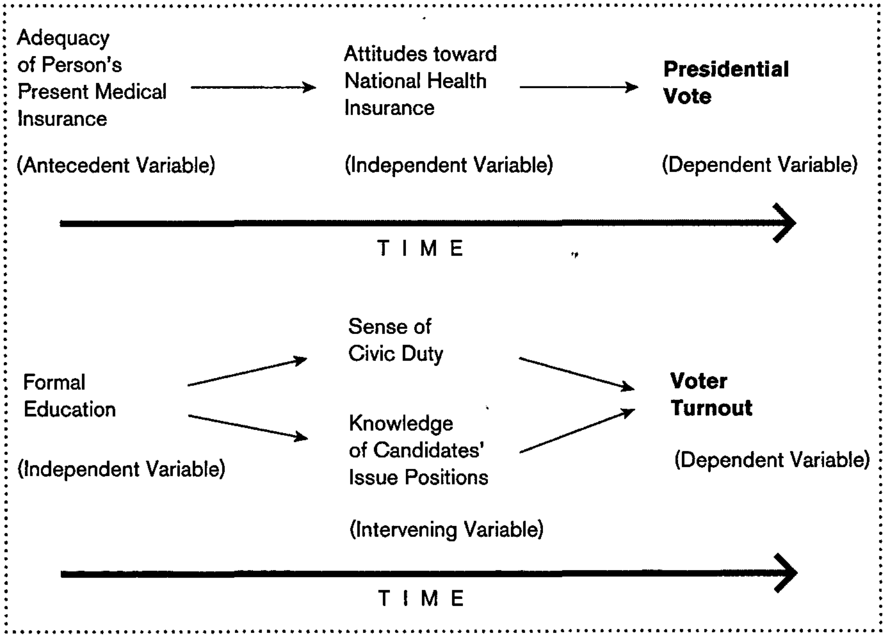
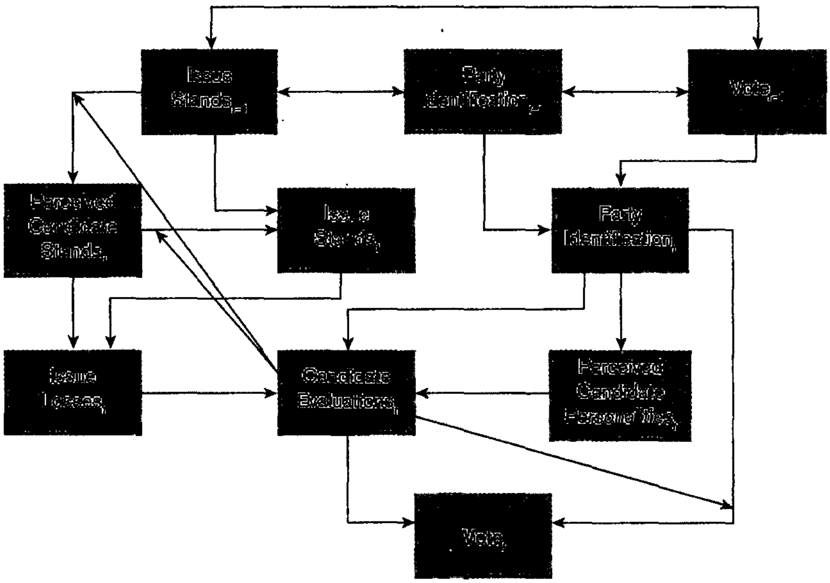
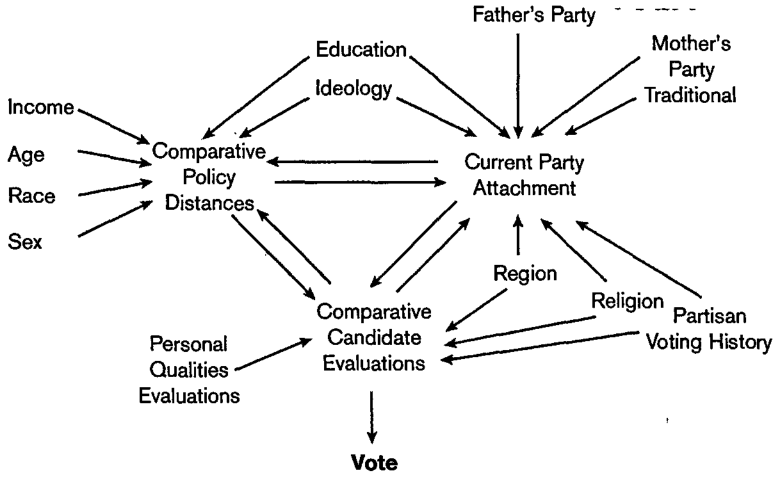


FIGURE 4-2 Two Causal Models of Vote Choice


Issue Stands$_{t-1}$
Party Identification$_{t-1}$
Vote$_{t-1}$
Perceived Candidate Stands$_t$
Issue Stands$_t$
Party Identification$_t$
Issue Losses$_t$
Candidate Evaluations$_t$
Perceived Candidate Personalities$_t$
Vote$_t$

Source: Gregory B. Markus and Philip E. Converse, "A Dynamic Simultaneous Equation Model of Electoral Choice," *American Political Science Review* 73 (December 1979): 1059. Copyright © 1979 American Political Science Association. Reprinted with permission of Cambridge University Press.


Father's Party
Education
Mother's Party Traditional
Ideology
Income
Age
Race
Sex
Comparative Policy Distances
Current Party Attachment
Region
Religion
Partisan Voting History
Comparative Candidate Evaluations
Personal Qualities Evaluations
Vote

Source: Benjamin I. Page and Calvin C. Jones, "Reciprocal Effects of Policy Preferences, Party Loyalties and the Vote," *American Political Science Review* 73 (December 1979): 1083. Copyright © 1979 American Political Science Association. Reprinted with permission of Cambridge University Press.

Characteristics of Good Hypotheses

For a hypothesis to be tested adequately and persuasively, it must be stated properly. It is important to start a research project with a clearly stated hypothesis because it provides the foundation for subsequent decisions and steps in the research process. A poorly formulated hypothesis often indicates confusion about the relationship to be tested or can lead to mistakes that will limit the value or the meaning of any findings. Many students find it challenging to write a hypothesis that precisely states the relationship to be tested: it takes practice to write consistently well-worded hypotheses. A good hypothesis has six characteristics: (1) it is an empirical statement, (2) it is stated as a generality, (3) it is plausible, (4) it is specific, (5) it is stated in a manner that corresponds to the way in which the researcher intends to test it, and (6) it is testable.

EMPIRICAL STATEMENT. Hypotheses should be empirical, rather than normative statements. Consider someone who is interested in democracy. If the researcher hypothesizes that "Democracy is the best form of government," he or she has formulated a normative, nonempirical statement that cannot be tested. The statement communicates the preference of the researcher; it does not explain a phenomenon. Instead, this researcher ought to be able to state how the central concept—in this case, democracy—is related to other concepts (such as literacy, size of population, geographical isolation, and economic development). Therefore, to produce an acceptable hypothesis, the researcher ought to make an educated guess about the relationship between democracy and another of these concepts; for example, "Democracy is more likely to be found in countries with high literacy than in countries with low literacy." This hypothesis now proposes a relationship between two phenomena that can be observed empirically. Or one might think that democracy is preferable to other systems because it produces higher standards of living. We cannot prove that one thing is preferable to another, but we could certainly compare countries on numerous measures of well-being, such as health status. The conclusion might then be "Compared with people living under dictatorships, citizens of democracies have higher life expectancies." Whether the hypothesis is confirmed empirically is not necessarily related to whether the researcher thinks the phenomenon (in this case, democracy) is good or bad.

To be sure, empirical knowledge can be relevant for normative inquiry. Often, people reach normative conclusions based on their evaluation of empirical relationships. Someone might reason, for example, that negative campaign ads cause voters to become disgusted with politics and not vote in elections; one might further reason that because low turnout is bad, negative campaign ads are bad as well. The first part of the assertion is an empirical statement, which could be investigated using the techniques developed in this book, whereas the next two (low turnout and negative ads being bad) are normative statements.

GENERALITY. A second characteristic of a good hypothesis is generality. It should propose a relationship pertaining to many occurrences of a phenomenon rather than just to one occurrence. Knowledge about the causes of particular occurrences of a phenomenon could be helpful in formulating more general guesses about the relationships between concepts, but with a general hypothesis, we attempt to expand the scope of our knowledge beyond individual cases. Stating hypotheses in the plural form, rather than the singular, makes it clear that testing the hypothesis will involve more than one case.

The four hypotheses in the left column below are too narrow, whereas the four hypotheses in the right column are more general and more acceptable as research propositions:

Senator X voted for a bill because it is the president's bill and they both are Democrats.	Senators are more likely to vote for bills sponsored by the president if they belong to the same political party as the president.
The United States is a democracy because its population is affluent.	Countries with high levels of affluence are more likely to be democracies than countries with low levels of affluence.
The United States has more murders than other countries because so many people own guns there.	Countries with more guns per capita will experience more murders per capita than countries with fewer guns.
Joe is a liberal because his mother is one too.	People tend to adopt political viewpoints similar to those of their parents.

Note that in each of the hypotheses on the right, the concepts being related in the hypothesis become clearer, as does the general nature of the relationship. So, for example, senators' support or opposition to bills sponsored by a president is thought to be influenced by whether or not they belong to the same party as the president. This hypothesis would apply to both Democratic and Republican senators.

PLAUSIBILITY. A third characteristic of a good hypothesis is that it be plausible. There should be some logical reason for thinking that it might be confirmed. Of course, since a hypothesis is a guess about a relationship, whether it will be confirmed cannot be known for certain. Any number of hypotheses could be thought of and tested, but many fewer are plausible ones. For example, if a researcher hypothesized that "people who eat dry cereal for breakfast are more likely to be liberal than are people who eat eggs," we would question his or her logic even though the form of the hypothesis may be perfectly acceptable. It is difficult to imagine why this hypothesis would be confirmed.

A researcher should therefore be able to justify why the relationship in each hypothesis is plausible and could be supported. The need to formulate plausible hypotheses is one of the reasons why researchers conduct a literature review early in their research projects. Literature reviews can acquaint researchers with both general theories and specific hypotheses that have been advanced by others. There are no hard and fast rules to ensure plausibility, however. After all, people used to think that "germs cause diseases" was an implausible hypothesis and that "dirt may be turned into gold" was a plausible one.

SPECIFICITY. The fourth characteristic of a good hypothesis is that it is specific. The researcher should not simply state that variables are associated; rather, he or she should indicate the direction of the expected relationship between two or more variables. Following are examples of **directional hypotheses** that specify the nature of the relationship between concepts:

- Median family income is higher in urban counties than in rural counties.
- States that are characterized by a "moralistic" political culture will have higher levels of voter turnout than will states with an "individualistic" or "traditionalistic" political culture.

The first hypothesis indicates which relative values of median family income are related to which type or category of county. Similarly, the second hypothesis predicts a particular relationship between specific types of political culture (the independent variable) and voter turnout (the dependent variable).

The direction of the relationship between concepts is referred to as a **positive relationship** if the concepts are predicted to increase in size together or decrease in size together; that is, as *X* increases, so does *Y*, and as *X* decreases, so does *Y*. The following are examples of hypotheses that predict positive relationships:

- The more education a person has, the higher his or her income.
- As the percentage of a country's population that is literate increases, the country's political process becomes more democratic.
- The older people become, the more likely they are to be conservative.
- People who read a daily newspaper are more informed about current events than are people who do not read a daily newspaper.
- The lower a state's per capita income, the less money the state spends per pupil on education.

If, however, the researcher thinks that as one concept increases in size or amount, another one will decrease in size or amount, then a **negative relationship** is suggested, as in the following examples:

- Older people are less tolerant of social protest than are younger people.
- The more income a person has, the less concerned about mass transit the person will be.
- More affluent countries have less property crime than do poorer countries.

In addition, the concepts used in a hypothesis should be defined carefully. For example, a hypothesis that suggests "There is a relationship between personality and political attitudes" is far too ambiguous. What is meant by personality? Which political attitudes? A more specific reformulation of this hypothesis might be "The more self-esteem a person has, the less likely the person is to be an isolationist." Now personality has been narrowed to self-esteem, and the political attitude has been defined as isolationism—both more precise concepts, although not precise enough. Eventually, even these two terms must be given more precise definitions when it comes to measuring them. (We return to the topic of defining concepts later in this chapter and further discuss the challenge of measuring concepts in chapter 5.) As the concepts become more clearly defined, the researcher is better able to specify the direction of the hypothesized relationship.

Following are four examples of ambiguous hypotheses that have been made more specific:

How a person votes for president depends on the information he or she is exposed to.	The more information favoring candidate X a person is exposed to during a political campaign, the more likely that person is to vote for candidate X.
A country's geographical location matters for the type of political system it develops.	The more borders a country shares with other countries, the more likely that country is to have a nondemocratic political process.
A person's capabilities affect his or her political attitudes.	The more intelligent a person is, the more likely he or she is to support civil liberties.
Guns do not cause crime.	People who own guns are less likely to be the victims of crimes than are persons who do not own guns.

CORRESPONDENCE TO THE WAY IN WHICH THE RESEARCHER INTENDS TO TEST THE HYPOTHESIS. A fifth characteristic of a good hypothesis is that it is stated in a manner that corresponds to the way in which the researcher intends to test it—that is, it should be "consistent with the data."[8] For example, although the hypothesis "Higher levels of literacy are associated with higher levels of democracy" does state how the concepts are related, it does not indicate how the researcher plans to test the hypothesis. In contrast, the hypothesis "As the percentage of a country's population that is literate increases, the country's political process becomes more democratic" suggests that the researcher is proposing to use a time series design by measuring the literacy rate and the amount of democracy for a country or countries at several different times to see if increases in democracy are associated with increases in literacy (that is, if changes in one concept lead to changes in another).

If, however, the researcher plans to test the hypothesis by measuring the literacy rates and levels of democracy for many countries at one point in time to see if those with higher literacy rates also have higher levels of democracy, it would be better to rephrase the hypothesis as "Countries with higher literacy rates tend to be more democratic than countries with lower literacy rates." This way of phrasing the hypothesis reflects that the researcher is planning to use a cross-sectional research design to compare the levels of democracy in countries with different literacy rates. This differs from comparing a country's level of democracy at more than one point in time to see if it changes in concert with changes in literacy.

TESTABILITY. Finally, a good hypothesis is testable. It must be possible and feasible to obtain data that will allow one to test the hypothesis. Hypotheses for which either confirming or disconfirming evidence is impossible to gather are not subject to testing, and hence are unusable for empirical purposes.

Consider this example of a promising yet difficult-to-test hypothesis: "The more a child is supportive of political authorities, the less likely that child will be to engage in political dissent as an adult." This hypothesis is general, plausible, fairly specific, and empirical, but in its current form it cannot be tested because, to our knowledge, no data exist to verify the proposition. The hypothesis requires data that measure a set of attitudes for individuals when they are children and a set of behaviors when they are adults. Consequently, a frustrating practical barrier prevents the testing of an otherwise acceptable hypothesis. Students in one-semester college courses on research methods often run up against practical constraints. A semester is not usually long enough to collect and analyze data, and some data

[8] This term is used by Susan Ann Kay in *Introduction to the Analysis of Political Data* (Englewood Cliffs, N.J.: Prentice Hall, 1991), 6.

may be too expensive to acquire. In fact, many interesting hypotheses go untested simply because even professional researchers do not have the resources to collect the data necessary to test them.

Hypotheses stated in tautological form are also untestable. A **tautology** is a statement linking two concepts that mean essentially the same thing; for example, "The less support there is for a country's political institutions, the more tenuous the stability of that country's political system." This hypothesis would be difficult to disconfirm because the two concepts—support for political institutions and stability of a political system—are so similar. To conduct a fair test, one would have to measure independently—in different ways—the support for the political institutions and the stability of the political system.

In their study of government maltreatment of citizens, Steven C. Poe and C. Neal Tate defined human rights abuses as coercive activities (such as murder, torture, forced disappearance, and imprisonment of persons for their political views) designed to induce compliance.[9] Other researchers have included lack of democratic processes and poor economic conditions in their definitions of human rights abuses, but Poe and Tate did not include these concepts because they wanted to use democratic rights and economic conditions as independent variables explaining variation in human rights abuses by governments.

Many hypotheses, then, are not formulated in a way that permits an informative test of them with empirical research. Readers of empirical research in political science, as well as researchers themselves, should take care that research hypotheses are empirical, general, plausible, specific, consistent with the data, and testable. Hypotheses that do not share these characteristics are likely to cause difficulty for the researcher and reader alike and make a minimal contribution to scientific knowledge.

Specifying Units of Analysis

In addition to proposing a relationship between two or more variables, a hypothesis also specifies, or strongly implies, the types or levels of political actor to which the hypothesis is thought to apply. This is called the **unit of analysis** of the hypothesis, and it also must be selected thoughtfully. A clearly established unit of analysis structures and helps to organize the collection of data to measure variables of interest.

As noted in chapter 2, political scientists are interested in understanding the behavior or properties of all sorts of political actors (individuals, groups, states,

[9] Steven C. Poe and C. Neal Tate, "Repression of Human Rights to Personal Integrity in the 1980s: A Global Analysis," *American Political Science Review* 88, no. 4 (1994): 853–72.

government agencies, organizations, regions, nations) and events (elections, wars, conflicts). The particular type of actor whose political behavior is named in a hypothesis is the unit of analysis for the research project. In a legislative behavior study, for example, the individual members of the House of Representatives might be the units of analysis in the following hypothesis:

- Members of the House who belong to the same party as the president are more likely to vote for legislation desired by the president than are members who belong to a different party.

In the following hypothesis, a city is the unit of analysis, since attributes of cities are being explored:

- Northeastern cities are more likely to have mayors, while western cities have city managers.

Civil wars are the units of analysis in this hypothesis:

- Civil wars that are halted by negotiated peace agreements are less likely to re-erupt than are those that cease due to the military superiority of one of the parties to the conflict.

Elections are the unit of analysis in this example:

- Elections in which the contestants spend the same amount of money tend to be decided by closer margins of victory than elections in which one candidate spends a lot more than the other candidate(s).

Finally, consider this proposition:

- The more affluent a country is, the more likely it is to have democratic political institutions.

Here the unit of analysis is the country. It is the measurement of national characteristics—affluence (the independent variable) and democratic political institutions (the dependent variable)—that is relevant to testing this hypothesis. In sum, the research hypothesis indicates the researcher's unit of analysis and the behavior or attributes that must be measured for that unit.

Cross-Level Analysis: Ecological Inference and Ecological Fallacy

Sometimes researchers conduct what is called **cross-level analysis**. In this type of analysis, researchers use data collected for one unit of analysis to make inferences about another unit of analysis. Christopher H. Achen and W. Phillips

Shively pointed out that "for reasons of cost or availability, theories and descriptions referring to one level of aggregation are frequently testable only with data from another level."[10] A discrepancy between the unit of analysis specified in a hypothesis and the entities whose behavior is empirically observed can cause problems, however.

A frequent goal of cross-level analysis is to make an **ecological inference**—that is, to use aggregate data to study the behavior of individuals.[11] Data of many kinds are collected for school districts, voting districts, counties, states, nations, or other aggregates in order to make inferences about individuals. The relationships between schools' average test scores and the percentage of children receiving subsidized lunches, national poverty and child mortality rates, air pollution indexes and the incidence of disease in cities, and the severity of state criminal penalties and crime rates are examples of relationships explored using aggregate data. The underlying hypotheses of such studies are that children who receive subsidized lunches score lower on standardized tests, that poor children are more likely to die of childhood diseases, that individuals' health problems are due to their exposure to air pollutants, and that harsh penalties deter individuals from committing crimes. Yet, if a relationship is found between group indicators or characteristics, it does not necessarily mean that a relationship exists between the characteristics for individuals in the group. The use of information that shows a relationship for groups to infer that the same relationship exists for individuals when in fact there is no such relationship at the individual level is called an **ecological fallacy**.

Let's take a look at an example to see how an ecological fallacy might be committed as a result of failing to be clear about the unit of analysis. Suppose a researcher wants to test the hypothesis "Democrats are more likely to support a sales tax increase than are Republicans." Individuals are the unit of analysis in this hypothesis. If the researcher selects an election in which a sales tax increase was at issue and obtains the voting returns as well as data on the proportions of Democrats and Republicans in each election precinct, the data are aggregate data, not data on individual voters. If it is found that sales tax increases received more votes in precincts with a higher proportion of Democrats than in the precincts with a higher proportion of Republicans, the researcher might take this as evidence in support of the hypothesis. There is a fundamental problem with this conclusion, however. Unless a district is 100 percent Democratic or 100 percent Republican, the researcher cannot necessarily draw such a conclusion about the behavior of individuals from the behavior of election districts. It could be that support for a sales tax increase in a district with a high proportion of Democratic voters came mostly from non-Democrats and that most of the support

[10] Christopher H. Achen and W. Phillips Shively, *Cross-Level Inference* (Chicago: University of Chicago Press, 1995), 4.

[11] Ibid.

for a sales tax increase in the Republican districts came from Republicans. If this is the case, then the researcher would have committed an ecological fallacy. What was true at the aggregate level was not true at the individual level.

Let us take two hypothetical election precincts to illustrate how this fallacy could occur. Suppose we have Precinct 1, classified as a "Democratic" district, and Precinct 2, a "Republican" district. If the Democratic district voted 67 percent to 33 percent in favor of the sales tax increase, and the Republican district voted 53 percent to 47 percent in favor of the sales tax increase, we might be tempted to conclude that Democrats as individuals voted more heavily for the sales tax increase than did Republicans.

But imagine we peek inside each of the election precincts to see how individuals with different party affiliations behaved. Suppose we obtain information about individuals within the districts. The data in table 4-1 show that in the Democratic district, Democrats split 25–25 for the tax increase, Republicans voted 18–2 for it, and others voted 24–6 for it. This resulted in the 67–33 percent edge for the tax increase in Precinct 1. In the Republican district, Precinct 2, Democrats voted 16–24 against the tax increase, Republicans split 30–20 for it, and others voted 7–3 in favor. This resulted in the 53–47 percent margin for the sales tax increase in Precinct 2. When we compare the percentage of Democrats, Republicans, and others voting for the sales tax increase, the difference in the voting behavior of party identifiers becomes clearer. In both precincts, the percentage of Democrats voting for the sales tax increase was lower than that of the two other groups of voters. Fifty percent of the Democrats in Precinct 1 voted for the sales tax increase, compared with 90 percent of the Republican voters and 80 percent of the others. In Precinct 2, only 40 percent of the Democrats voted for the sales tax increase, compared with 60 percent of the Republicans and 70 percent of the other voters. In other words, Republicans as individuals were more likely to have voted for the tax increase than were Democrats as individuals in both precincts. Knowing only the precinct-level totals gave the opposite impression. When the results for both districts are combined and broken down by party, we see that, overall, 68.6 percent of Republicans and 45.6 percent of Democrats voted for the sales tax increase.

In the research by Ansolabehere and his colleagues discussed in chapter 1, the tone of campaign advertising and the roll-off rates were measured in thirty-four Senate races, and states with races characterized by a negative tone had higher roll-off rates than states with positive campaigns.[12] The inference is that those individuals exposed to negative campaign ads are less likely to vote than are those exposed to positive campaign ads. But the researchers lacked data that showed the relationship

[12] Stephen D. Ansolabehere, Shanto Iyengar, Adam Simon, and Nicholas Valentino, "Does Attack Advertising Demobilize the Electorate?" *American Political Science Review* 88, no. 4 (1994): 829–38. Available at http://weber.ucsd.edu/~tkousser/Ansolabehere.pdf

TABLE 4-1 Voting by Democrats, Republicans, and Others for a Sales Tax Increase

	Raw Vote			Percentage of Vote	
Party ID	Number	Against tax increase	For tax increase	Against tax increase	For tax increase
Precinct 1					
Democrats	50	25	25	50.0	50.0
Republicans	20	2	18	10.0	90.0
Other	30	6	24	20.0	80.0
Total	100	33	67	33.0	67.0
Precinct 2					
Democrats	40	24	16	60.0	40.0
Republicans	50	20	30	40.0	60.0
Other	10	3	7	30.0	70.0
Total	100	47	53	47.0	53.0
Voting of individuals					
Democrats	90	49	41	54.4	45.6
Republicans	70	22	48	31.4	68.6
Other	40	9	31	22.5	77.5
Total	200	80	120	40.0	60.0

Note: Hypothetical data.

between actual exposure to campaign ads of individuals and their voting behavior in the Senate elections. Remember, however, that the researchers examined and reported on individual-level data obtained from experiments, so they did not rely just on aggregate data to test their hypotheses about individuals.

Use of aggregate data to examine hypotheses that pertain to individuals may be unavoidable in some situations because individual-level data are lacking. Achen and Shively pointed out that before the development of survey research, aggregate data generally were the only data available and were used routinely by political scientists.[13] Several statistical methods have been developed to try to adjust inferences from aggregate-level data, although a discussion of these is beyond the scope of this book.[14]

Another mistake researchers sometimes make is to mix different units of analysis in the same hypothesis. "The more education a person has, the more democratic his or her country is" doesn't make much sense because it mixes the individual and country as units of analysis. However, though "The smaller a government agency, the happier its workers" concerns an attribute of an agency and an attribute of individuals, it does so in a way that makes sense. The size of the agency in which individuals work may be an important aspect of the context or environment in which the individual phenomenon occurs and may influence the individual attribute. In this case, the unit of analysis is clearly the individual, but a phenomenon that is experienced by many cases is used to explain the behavior of individuals, some of whom may well be identically situated.

In short, a researcher must be careful about the unit of analysis specified in a hypothesis and its correspondence with the unit measured. In general, a researcher should not mix units of analysis within a hypothesis.

Defining Concepts

Political scientists are interested in why people or social groupings (organizations, political parties, legislatures, states, countries) behave in a certain way or have particular attributes or properties. The words that we choose to describe these behaviors or attributes are called *concepts*. Concepts should be accurate, precise, and informative. Clear definitions of the concepts of interest to us are important if we are to develop specific hypotheses and avoid tautologies. Clear definitions also are important so that the knowledge we acquire from testing our hypotheses is transmissible and empirical.

In our daily life, we use concepts frequently to name and describe features of our environment. For example, we describe some snakes as poisonous and others as nonpoisonous, some politicians as liberal and others as conservative, some friends as shy and others as extroverted. These attributes, or concepts, are useful to us because they help us observe and understand aspects of our environment, and they help us communicate with others.

[13] Achen and Shively, *Cross-Level Inference*, 5–10.

[14] For example, see Gary King, *A Solution to the Ecological Inference Problem* (Princeton, N.J.: Princeton University Press, 1997); Achen and Shively, *Cross-Level Inference;* and Barry C. Burden and David C. Kimball, "Measuring Ticket Splitting," chap. 3 in *Why Americans Split Their Tickets: Campaigns, Competition, and Divided Government* (Ann Arbor: University of Michigan Press, 2002).

Concepts also contribute to the identification and delineation of the scientific disciplines within which research is conducted. In fact, to a large extent a discipline maintains its identity because different researchers within it share a concern for the same concepts. Physics, for example, is concerned with the concepts of gravity and mass (among others); sociology, with social class and social mobility; psychology, with personality and deviance. By contrast, political science is concerned with concepts such as democracy, power, representation, justice, and equality. The boundaries of disciplines are not well defined or rigid, however. Political scientists, developmental psychologists, sociologists, and anthropologists all share an interest in how new members of a society are socialized into the norms and beliefs of that society, for example. Nonetheless, because a particular discipline has some minimal level of shared consensus concerning its significant concepts, researchers can usually communicate more readily with other researchers in the same discipline than with researchers in other disciplines.

A shared consensus over those concepts thought to be significant is related directly to the development of theories. Thus, a theory of politics will identify significant concepts and suggest why they are central to an understanding of political phenomena. Concepts are developed through a process by which some human group (tribe, nation, culture, profession) agrees to give a phenomenon or a property a particular name. The process is ongoing and somewhat arbitrary and does not ensure that all peoples everywhere will give the same phenomena the same names. In some areas of the United States, for example, a *soda* is a carbonated beverage, while in other areas it is a drink with ice cream in it. Likewise, the English language has only one word for *love*, whereas the Greeks have three words to distinguish among romantic love, familial love, and generalized feelings of affection.[15] Concepts disappear from a group's language when they are no longer needed, and new ones are invented as new phenomena are noticed that require names (for example, computer *programs* and *software, cultural imperialism,* and *hyperkinetic* behavior).

Some concepts—such as *car, chair,* and *vote*—are fairly precise because there is considerable agreement about their meaning. Others are more abstract and lend themselves to differing definitions—for example, *liberalism, crime, democracy, equal opportunity, human rights, social mobility,* and *alienation.* A similar concept is *orange.* Although there is considerable agreement about it (orange is not usually confused with purple), the agreement is less than total (whether a particular object is orange or red is not always clear).

Many interesting concepts that political scientists deal with are abstract and lack a completely precise, shared meaning. This hinders communication concerning research and creates uncertainty regarding the measurement of a phenomenon.

[15] Kenneth R. Hoover, *The Elements of Social Scientific Thinking* (New York: St. Martin's, 1980), 18–19.

Consequently, a researcher must explain what is meant by the concept so that a measurement strategy may be developed and so that those reading and evaluating the research can decide if the meaning accords with their own understanding of the term. Although some concepts that political scientists use—such as *amount of formal education, presidential vote,* and *amount of foreign trade*—are not particularly abstract, other concepts—such as *partisan realignment, political integration,* and *regime support*—are far more abstract and need more careful consideration and definition.

Suppose, for example, that a researcher is interested in the kinds of political systems that different countries have and, in particular, why some countries are more democratic than others. *Democracy* is consequently a key concept that needs definition and measurement. The word contains meaning for most of us; that is, we have some idea of what is democratic and what is not. But once we begin thinking about the concept, we quickly realize that it is not as clear as we originally thought. In fact, a group of researchers wrote in 2011, "Perhaps no other concept is as central to policymakers and scholars. Yet, there is no consensus about how to conceptualize and measure regimes such that meaningful comparisons can be made through time and across countries."[16] To some, a country is democratic if it has "competing political parties, operating in free elections, with some reasonable level of popular participation in the process."[17] To others, a country is democratic only if legal guarantees protect free speech, the press, religion, and the like. To others, a country is democratic if the political leaders make decisions that are acceptable to the populace. And to still others, democracy implies equality of economic opportunity among the citizenry. If a country has all these attributes, it would be called a democracy by any of the criteria, and there would be no problem classifying the country. But if a country possesses only one of these attributes, its classification would be uncertain, since by some definitions it would be democratic but by others it would not. Different definitions require different measurements and may result in different research findings. Hence, defining one's concepts is important, particularly when the concept is so abstract as to make shared agreement difficult.

Concept definitions have a direct impact on the quality of knowledge produced by research studies. Suppose, for example, that a researcher is interested in the connection between economic development and democracy, the working hypothesis being that countries with a high level of economic development will be more likely to have democratic forms of government. And suppose that there are two definitions of *economic development* and two definitions of *democracy* that might be used in the research. Finally, suppose that the researcher has data on twelve

[16] Michael Coppedge and John Gerring, "Conceptualizing and Measuring Democracy: A New Approach," *Perspectives on Politics* 9, no. 2 (2011): 247–67.

[17] W. Phillips Shively, *The Craft of Political Research* (Englewood Cliffs, N.J.: Prentice Hall, 1980), 33.

TABLE 4-2 **Concept Development: The Relationship between Economic Development and Democracy**

Is the country economically developed?			
		By definition 1	
		Yes	No
By definition 2	Yes	A,B,C	G,H,I
	No	D,E,F	J,K,L

Is the country a democracy?			
		By definition 1	
		Yes	No
By definition 2	Yes	D,E,F	J,K,L
	No	A,B,C	G,H,I

countries (A–L) included in the study. In table 4-2, we show that the definition selected for each concept has a direct bearing on how different countries are categorized on each attribute. By definition 1, countries A, B, C, D, E, and F are economically developed; however, by definition 2, countries A, B, C, G, H, and I are. By definition 1, countries A, B, C, D, E, and F are democracies; by definition 2, countries D, E, F, J, K, and L are.

This is only the beginning of our troubles, however. When we look for a pattern involving the economic development and democracy of countries, we find that our answer depends mightily on how we have defined the two concepts. If we use the first definitions of the two concepts, we find that all economically developed countries are also democracies (A, B, C, D, E, F), which supports our hypothesis. If we use the first definition for economic development and the second for democracy (or vice versa), half of the economically developed nations are democracies and half are not. If we use the second definitions of both concepts, none of the economically developed countries is a democracy, whereas all of the undeveloped countries are (D, E, F, J, K, L). In other words, because of our inability to formulate a precise definition of the two concepts, and because the two definitions of each concept yield quite different categorizations of the twelve countries, our hypothesis could be either confirmed or disconfirmed by the data at hand. Our conceptual confusion has put us in a difficult position.

Consider another example. Suppose a researcher is interested in why some people are liberal and some are not. In this case, we need to define what is meant by *liberal* so that those who are liberal can be identified. *Liberal* is a frequently used term, but it has many different meanings: one who favors change, one who favors redistributive income or social welfare policies, one who favors increased government spending and taxation, or one who opposes government interference in the political activities of its citizens. If a person possesses all these attributes, there is no problem deciding whether or not he or she is a liberal. A problem arises, however, when a person possesses some of these attributes but not others.

The examples here illustrate the elusive nature of concepts and the need to define them. The empirical researcher's responsibility to define terms is a necessary and challenging one. Unfortunately, many of the concepts used by political science researchers are abstract and require careful thought and extensive elaboration.

Researchers can clarify the concept definitions they use simply by making the meanings of key concepts explicit. This requires researchers to think carefully about the concepts used in their research and to share their meanings with others. Other researchers often challenge concept definitions, requiring researchers to elaborate upon and justify their meanings.

Another way in which researchers get help defining concepts is by reviewing and borrowing (possibly with modification) definitions developed by others in the field. For example, a researcher interested in the political attitudes and behavior of the American public would find the following definitions of key concepts in the existing literature:

- *Political participation:* "Those activities by private citizens that are more or less directly aimed at influencing the selection of government personnel and/or the actions they take"[18]
- *Political violence:* "All collective attacks within a political community against the political regime, its actors—including competing political groups as well as incumbents—or its policies"[19]
- *Political efficacy:* "The feeling that individual political action does have, or can have, an impact upon the political processes—that it is worthwhile to perform one's civic duties"[20]
- *Belief system:* "A configuration of ideas and attitudes in which the elements are bound together by some form of constraint or functional interdependence"[21]

Each of these concepts is somewhat vague and lacks complete shared agreement about its meaning. Furthermore, it is possible to raise questions about each of these concept definitions. Notice, for example, that the definition of *political participation* excludes the possibility that government employees (presumably "nonprivate" citizens) engage in political activities and that the definition of *political efficacy* excludes the impact of collective political action on political processes. Consequently, we may find these and other concept definitions inadequate and revise them to capture more accurately what we mean by the terms.

Over time, a discipline cannot proceed very far unless some minimal agreement is reached about the meanings of the concepts with which scientific research is concerned. Researchers must take care to think about the phenomena named in a research project and make explicit the meanings of any problematic concepts.

[18] Sidney Verba and Norman H. Nie, *Participation in America: Political Democracy and Social Equality* (New York: Harper and Row, 1972), 2.

[19] Ted Robert Gurr, *Why Men Rebel* (Princeton, N.J.: Princeton University Press, 1970), 3–4.

[20] Angus Campbell, Gerald Gurin, and Warren E. Miller, *The Voter Decides* (Evanston, Ill.: Row, Peterson, 1954), 187.

[21] Philip E. Converse, "The Nature of Belief Systems in Mass Publics," in *Ideology and Discontent*, ed. David E. Apter (New York: Free Press, 1964), 207.

Conclusion

In this chapter, we discussed the beginning stages of a scientific research project. A research project must provide—to both the producer and the consumer of social scientific knowledge—the answers to these important questions: What phenomenon is the researcher trying to understand and explain? What explanation has the researcher proposed for the political behavior or attributes in question? What are the meanings of the concepts used in this explanation? What specific hypothesis relating two or more variables will be tested? What is the unit of analysis for the observations? If these questions are answered adequately, then the research will have a firm foundation.

TERMS INTRODUCED

Antecedent variable. An independent variable that precedes other independent variables in time.

Arrow diagram. A pictorial representation of a researcher's explanatory scheme.

Constant. A concept or variable whose values do not vary.

Cross-level analysis. The use of data at one level of aggregation to make inferences at another level of aggregation.

Dependent variable. The phenomenon thought to be influenced, affected, or caused by some other phenomenon.

Directional hypothesis. A hypothesis that specifies the expected relationship between two or more variables.

Ecological fallacy. The fallacy of deducing a false relationship between the attributes or behavior of individuals based on observing that relationship for groups to which the individuals belong.

Ecological inference. The process of inferring a relationship between characteristics of individuals based on group or aggregate data.

Hypothesis. A tentative or provisional or unconfirmed statement that can (in principle) be verified.

Independent variable. The phenomenon thought to influence, affect, or cause some other phenomenon.

Intervening variable. A variable coming between an independent variable and a dependent variable in an explanatory scheme.

Negative relationship. A relationship in which the values of one variable increase as the values of another variable decrease.

Positive relationship. A relationship in which the values of one variable increase (or decrease) as the values of another variable increase (or decrease).

Tautology. A hypothesis in which the independent and dependent variables are identical, making it impossible to disconfirm.

Unit of analysis. The type of actor (individual, group, institution, nation) specified in a researcher's hypothesis.

SUGGESTED READINGS

Achen, Christopher H., and W. Phillips Shively. *Cross-Level Inference.* Chicago: University of Chicago Press, 1995.

King, Gary. *A Solution to the Ecological Inference Problem.* Princeton, N.J.: Princeton University Press, 1997.

King, Gary, Ori Rosen, and Martin A. Tanner, eds. *Ecological Inference: New Methodological Strategies.* Cambridge, UK: Cambridge University Press, 2004.

Outhwaite, William, and Stephen P. Turner, eds. *The Sage Handbook of Social Science Methodology.* Los Angeles, Calif.: Sage, 2007.

CHAPTER 5

The Building Blocks of Social Scientific Research:

Measurement

CHAPTER OBJECTIVES

5.1 Discuss the importance of operationalization in hypothesis measurement.

5.2 Explain why measurements of political phenomena must correspond closely to the original meaning of a researcher's concepts.

5.3 Summarize the ways in which accurate measurements must be reliable and valid.

5.4 Describe different levels of measurement and their importance of measurement precision.

5.5 Identify different types of multi-item measures.

IN THE PREVIOUS CHAPTERS, WE DISCUSSED the beginning stages of political science research projects: the choice of research topics, the formulation of scientific explanations, the development of testable hypotheses, and the definition of concepts. In this chapter, we take the next step toward testing hypotheses empirically. Before testing hypotheses, we must understand some issues involving the **measurement** of the concepts we have decided to investigate and how we record systematic observations using numerals or scores to create variables that represent the concepts for analysis.

In chapter 2, we said that scientific knowledge is based on empirical research. In order to test empirically the accuracy and utility of a scientific explanation for a political phenomenon, we will have to observe and measure the presence of the concepts we are using to understand that phenomenon. Furthermore, if this test is to be adequate, our measurements of the political phenomenon must be as accurate and precise as possible. The process of measurement is important because it provides the bridge between our proposed explanations and the empirical world they are supposed to explain. How researchers measure their concepts can have a significant impact on their findings; differences in measurement can lead to totally different conclusions.

Lane Kenworthy and Jonas Pontusson's investigation of income inequality in affluent countries illustrates well the impact on research findings of how a concept is measured.[1] One way to measure income distribution is to look at the earnings of full-time-employed individuals and to compare the incomes of those at the top and the bottom of the earnings distribution. Kenworthy and Pontusson argued that it is more appropriate to compare the incomes of households than incomes of individuals. The unemployed are excluded from the calculations of individual earnings inequality, but households include the unemployed. Also, low-income workers disproportionately drop out of the employed labor force. Using working-age household income reflects changes in employment among household members. Kenworthy and Pontusson found that when individual income was used as a basis for measuring inequality, inequality had increased the most in the United States, New Zealand, and the United Kingdom, all liberal market economies. They further found that income inequality had increased significantly more in these countries than in Europe's social market economies and Japan. When household income was used, the data indicated that inequality had increased in all countries with the exception of the Netherlands.

Another example involves the measurement of turnout rates (discussed in chapter 1). Political scientists have investigated whether turnout rates in the United States have declined in recent decades.[2] The answer may depend on how the number of eligible voters is measured. Should it be the number of all citizens of voting age, or should this number be adjusted to take into account those who are not eligible to vote, or should the turnout rate be calculated using just the number of registered voters as the potential voting population?

The researchers discussed in chapter 1 measured a variety of political phenomena, some of which posed greater challenges than others. Milner, Poe, and Leblang wanted to measure

[1] Lane Kenworthy and Jonas Pontusson, "Rising Inequality and the Politics of Redistribution in Affluent Countries," *Perspectives on Politics* 3, no. 3 (2005): 449–71. Available at http://www.u.arizona.edu/~lkenwor/pop2005.pdf

[2] See Walter Dean Burnham, "The Turnout Problem," in *Elections American Style*, ed. A. James Reichley (Washington, D.C.: Brookings Institution, 1987), 97–133; Michael P. McDonald and Samuel L. Popkin, "The Myth of the Vanishing Voter," *American Political Science Review* 95, no. 4 (2001): 963–74. Available at http://elections.gmu.edu/APSR%20McDonald%20and _Popkin_2001.pdf

three different types of human rights: personal integrity or security rights, subsistence rights, and civil and political rights. Each of these types of rights has multiple dimensions. For example, civil and political rights consist of both civil liberties, such as freedom of speech, as well as economic liberties, including private property rights. Jeffrey A. Segal and Albert D. Cover measured both the political ideologies and the written opinions of US Supreme Court justices in cases involving civil rights and liberties.[3] Valerie J. Hoekstra measured people's opinions about issues connected to Supreme Court cases and their opinions about the Court.[4] Richard L. Hall and Kristina Miler wanted to measure oversight activity by members of Congress, the number of times they were contacted by lobbyists, and whether members of Congress and lobbyists were pro-regulation or antiregulation.[5] And Stephen Ansolabehere, Shanto Iyengar, Adam Simon, and Nicholas Valentino measured the intention to vote reported by study participants to see if it was affected by exposure to negative campaign advertising.[6] In each case, some political behavior or attribute was measured so that a scientific explanation could be tested. All of these researchers made important choices regarding their measurements.

Devising Measurement Strategies

As we pointed out in chapter 4, researchers must define the concepts they use in their hypotheses through conceptualization. They also must decide how to measure the presence, absence, or amount of these concepts in the real world. Political scientists refer to this process as **operationalization**, or providing an **operational definition** of their concepts. Operationalization is deciding how to record empirical observations of the occurrence of an attribute or a behavior using numerals or scores.

Let us consider, for example, a researcher trying to explain the existence of democracy in different nations. If the researcher were to hypothesize that higher rates of literacy make democracy more likely, then a definition of two concepts—literacy and democracy—would be necessary. The researcher could then develop a

[3] Jeffrey A. Segal and Albert D. Cover, "Ideological Values and the Votes of U.S. Supreme Court Justices," *American Political Science Review* 83, no. 2 (1989): 557–65. Available at http://www.uic.edu/classes/pols/pols200mm/Segal89.pdf

[4] Valerie J. Hoekstra, *Public Reaction to Supreme Court Decisions* (New York: Cambridge University Press, 2003).

[5] Richard C. Hall and Kristina Miler, "What Happens after the Alarm? Interest Group Subsidies to Legislative Overseers," *Journal of Politics* 70, no. 4 (2008): 990–1005.

[6] Stephen Ansolabehere, Shanto Iyengar, Adam Simon, and Nicholas Valentino, "Does Attack Advertising Demobilize the Electorate?" *American Political Science Review* 88, no. 4 (1994): 829–38. Available at http://weber.ucsd.edu/~tkousser/Ansolabehere.pdf

strategy, based on the definitions of the two concepts, for measuring the existence and amount of both attributes in nations.

Suppose *literacy* was defined as "the completion of six years of formal education" and *democracy* was defined as "a system of government in which public officials are selected in competitive elections." These definitions would then be used to develop operational definitions of the two concepts. These operational definitions would indicate what should be observed empirically to measure both literacy and democracy, and they would indicate specifically what data should be collected to test the researcher's hypothesis. In this example, the operational definition of *literacy* might be "those nations in which at least 50 percent of the population has had six years of formal education, as indicated in a publication of the United Nations," and the operational definition of *democracy* might be "those countries in which the second-place finisher in elections for the chief executive office has received at least 25 percent of the vote at least once in the past eight years."

When a researcher specifies a concept's operational definition, the concept's precise meaning in a particular research study becomes clear. In the preceding example, we now know exactly what the researcher means by *literacy* and *democracy.* Since different people often mean different things by the same concept, operational definitions are especially important. Someone might argue that defining literacy in terms of formal education ignores the possibility that people who complete six years of formal education might still be unable to read or write well. Similarly, it might be argued that defining democracy in terms of competitive elections ignores other important features of democracy, such as freedom of expression and citizen involvement in government activity. In addition, the operational definition of *competitive elections* is clearly debatable. Is the "competitiveness" of elections based on the number of competing candidates, the size of the margin of victory, or the number of consecutive victories by a single party in a series of elections? Unfortunately, operational definitions are seldom absolutely correct or absolutely incorrect; rather, they are evaluated according to how well they correspond to the concepts they are meant to measure.

It is useful to think of arriving at the operational definition as being the last stage in the process of defining a concept precisely. We often begin with an abstract concept (such as democracy), then attempt to define it in a meaningful way, and finally decide in specific terms how we are going to measure it. At the end of this process, we hope to attain a definition that is sensible, close to our meaning of the concept, and exact in what it tells us about how to go about measuring the concept.

Let us consider another example: imagine that a researcher is interested in why some individuals are more liberal than others. The concept of *liberalism* might be defined as "believing that government ought to pursue policies that provide benefits for the less well-off." The task, then, is to develop an operational definition that can be used to measure whether particular individuals are liberal or not. The following question from the General Social Survey might be used to operationalize the concept:

> 73A. Some people think that the government in Washington ought to reduce the income differences between the rich and the poor, perhaps by raising the taxes of wealthy families or by giving income assistance to the poor. Others think that the government should not concern itself with reducing this income difference between the rich and the poor.
>
> Here is a card with a scale from 1 to 7. Think of a score of 1 as meaning that the government ought to reduce the income differences between rich and poor, and a score of 7 as meaning that the government should not concern itself with reducing income differences. What score between 1 and 7 comes closest to the way you feel? (CIRCLE ONE)[7]

An abstract concept, liberalism has now been given an operational definition that can be used to measure the concept for individuals. This definition is also related to the original definition of the concept, and it indicates precisely what observations need to be made. It is not, however, the only operational definition possible. Others might suggest that questions regarding affirmative action, same-sex marriage, school vouchers, the death penalty, welfare benefits, and pornography could be used to measure liberalism.

The important thing is to think carefully about the operational definition you choose and to try to ensure that the definition coincides closely with the meaning of the original concept. How a concept is operationalized affects how generalizations are made and interpreted. For example, general statements about liberals or conservatives apply to liberals or conservatives only as they have been operationally defined, in this case by this one question regarding government involvement in reducing income differences. As a consumer of research, you should familiarize yourself with the operational definitions used by researchers so that you are better able to interpret and generalize research results.

Examples of Political Measurements: Getting to Operationalization

Let us take a closer look at some operational definitions used by the political science researchers referred to in chapter 1, as well as some others. To measure the strength of a legislator's intervention in air pollution regulations proposed by the

[7] Question wording for the variable EQWLTH from GSS 1998 *Codebook.* Available at http://www.thearda.com/Archive/Files/Codebooks/GSS1998_CB.asp

Environmental Protection Agency, Hall and Miler coded and counted the number of substantive comments made by legislators challenging or defending the agency's proposed air quality regulations during five oversight hearings held in Congress and during the public comment period.[8] Agencies are required to maintain a public docket that contains all the comments received during the comment period. Transcripts were available for each of the hearings. The researchers ended up with two variables: one was the number of supporting comments; the other was the number of comments in opposition to the proposed regulation. To measure constituency interests in each of the members' districts, they measured the number of manufacturing jobs in each district and created an index of air pollution based on district levels of PM 10 particulate matter and ground-level ozone (the pollutants addressed by the proposed regulations). Because Hall and Miler were interested in investigating whether lobbyists targeted their efforts toward members of Congress friendly toward the lobbyists' positions, they needed to measure the pro- or antienvironmental policy positions for each member of Congress, and this variable had to measure position *before* the oversight hearings and regulatory comment period. Fortunately for the researchers, the leaders of the health and environmental coalition had classified members in terms of their likely support for the rule prior to the lobbying period and were willing to share their ratings. These measures were based on legislators' previous voting record on health and environmental issues.

The research conducted by Segal and Cover on the behavior of US Supreme Court justices is a good example of an attempt to overcome a serious measurement problem to test a scientific hypothesis.[9] Recall that Segal and Cover were interested, as many others have been before them, in the extent to which the votes cast by Supreme Court justices were dependent on the justices' personal political attitudes. Measuring the justices' votes on the cases decided by the Supreme Court is no problem; the votes are public information. But measuring the personal political attitudes of judges, *independent of their votes,* is a problem (remember the discussion in chapter 4 on avoiding tautologies, or statements that link two concepts that mean essentially the same thing). Many of the judges whose behavior is of interest have died, and it is difficult to get living Supreme Court justices to reveal their political attitudes through personal interviews or questionnaires. Furthermore, one ideally would like a measure of attitudes that is comparable across many judges and that measures attitudes related to the cases decided by the Court.

Segal and Cover limited their inquiry to votes on civil liberties cases between 1953 and 1987, so they needed a measure of related political attitudes for the judges serving on the Supreme Court over that same period. They decided to infer the judges' attitudes from the newspaper editorials written about them in four major

[8] Hall and Miler, "What Happens after the Alarm?"

[9] Segal and Cover, "Ideological Values and the Votes of U.S. Supreme Court Justices."

daily newspapers from the time each justice was appointed by the president until the justice's confirmation vote by the Senate. They selected the editorials appearing in two liberal papers and in two conservative papers. Trained analysts read the editorials and coded each paragraph for whether it asserted that a justice designate was liberal, moderate, or conservative (or if the paragraph was inapplicable) regarding "support for the rights of defendants in criminal cases, women and racial minorities in equality cases, and the individual against the government in privacy and First Amendment cases."[10]

Because of practical barriers to ideal measurement, then, Segal and Cover had to rely on an indirect measure of judicial attitudes *as perceived by four newspapers* rather than on a measure of the attitudes themselves. Although this approach *may* have resulted in flawed measures, it also permitted the test of an interesting and important hypothesis about the behavior of Supreme Court justices that had not been tested previously. Without such measurements, the hypothesis could not have been tested.

Next, let us consider research conducted by Bradley and his colleagues on the relationship between party control of government and the distribution and redistribution of wealth.[11] The researchers relied on the Luxembourg Income Study (LIS) database, which provides cross-national income data over time in OECD (Organisation for Economic Co-operation and Development) countries.[12] They decided, however, to make adjustments to published LIS data on income inequality. That data included pensioners. Because some countries make comprehensive provisions for retirees, retirees in these countries make little provision on their own for retirement. Thus, many of these people would be counted as "poor" before any government transfers. Including pensioners would inflate the pretransfer poverty level as well as the extent of income transfer for these countries. Therefore, Bradley and his colleagues limited their analysis to households with a head aged twenty-five to fifty-nine (thus excluding the student-age population as well) and calculated their own measures of income inequality from the LIS data. They argued that their data would measure redistribution across income groups, not life-cycle redistributions of income, such as transfers to students and retired persons. *Income* was defined as income from wages and salaries, self-employment income, property income, and private pension income. The researchers also made adjustments for household size using an equivalence scale, which adjusts the number of persons in a household to an equivalent number of adults. The equivalence scale takes into account the economies of scale resulting from sharing household expenses.

[10] Ibid., 559.

[11] David Bradley, Evelyne Huber, Stephanie Moller, Francoise Nielsen, and John D. Stephens, "Distribution and Redistribution in Postindustrial Democracies," *World Politics* 55, no. 2 (2003): 193–228.

[12] For information on the LIS database, see http://www.lisdatacenter.org/our-data/lis-database/

Martin P. Wattenberg and Craig Leonard Brians measured exposure by responses to a survey question that asked respondents if they recalled a campaign ad and whether or not it was negative or positive in tone.[13] Finally, Ansolabehere and his colleagues measured exposure to negative campaign ads in the 1990 Senate elections by accessing newspaper and magazine articles about the campaigns and determining how the tone of the campaigns was described in these articles.[14]

The cases discussed here are good examples of researchers' attempts to measure important political phenomena (behaviors or attributes) in the real world. Whether the phenomenon in question was judges' political attitudes, income inequality, the tone of campaign advertising, or the attitudes and behavior of legislators, the researchers devised measurement strategies that could detect and measure the presence and amount of the concept in question. These observations were then generally used as the basis for an empirical test of the researchers' hypotheses.

To be useful in providing scientific explanations for political behavior, measurements of political phenomena must correspond closely to the original meaning of a researcher's concepts. They must also provide the researcher with enough information to make valuable comparisons and contrasts. Hence, the quality of measurements is judged in regard to both their *accuracy* and their *precision*.

The Accuracy of Measurements

Because we are going to use our measurements to test whether or not our explanations for political phenomena are valid, those measurements must be as accurate as possible. Inaccurate measurements may lead to erroneous conclusions, since they will interfere with our ability to observe the actual relationship between two or more variables.

There are two major threats to the accuracy of measurements. Measures may be inaccurate because they are *unreliable* and/or because they are *invalid.*

Reliability

Reliability describes the consistency of results from a procedure or measure in repeated tests or trials. In the context of measurement, a reliable measure is one

[13] Martin P. Wattenberg and Craig Leonard Brians, "Negative Campaign Advertising: Demobilizer or Mobilizer?" *American Political Science Review* 93, no. 4 (1999): 891–99. Available at http://weber.ucsd.edu/~tkousser/Wattenberg.pdf

[14] Ansolabehere, Iyengar, Simon, and Valentino, "Does Attack Advertising Demobilize the Electorate?"

that produces the same result each time the measure is used. An unreliable measure is one that produces inconsistent results—sometimes higher, sometimes lower.[15]

Suppose, for example, you want to measure support for the president among college students. You select two similar survey questions (Q1 and Q2) and ask the participants in a random sample of students to answer each question. The results from this sample were 50 percent support for the president using Q1 and 50 percent support for the president using Q2. But what might you find if you ask the same questions of multiple random samples of students? Will the results from each question remain consistent, assuming that the samples are identical? If a second sample of students is polled, you may find the same result, 50 percent, for Q1 but 60 percent for Q2. If you were to ask Q1 of multiple random samples of students and the result was consistently 50 percent, you could assert that your measure, Q1, is reliable. If Q2 were asked to multiple random samples of students and each sample of students returned different answers ranging somewhere between 40 percent and 60 percent, you could conclude that Q2 is less reliable than Q1 because Q2 generates inconsistent results each time it is used.

Likewise, you can assess the reliability of procedures as well. Suppose you are given the responsibility of counting a stack of one thousand paper ballots for some public office. The first time you count them, you obtain a particular result. But as you were counting the ballots, you might have been interrupted, two or more ballots might have stuck together, some might have been blown onto the floor, or you might have written down the totals incorrectly. As a precaution, then, you count them five more times and get four other people to count them once each as well. The similarity of the results of all ten counts would be an indication of the reliability of the counting process.

Similarly, suppose you wanted to test the hypothesis that the *New York Times* is more critical of the federal government than is the *Wall Street Journal.* This would require you to measure the level of criticism found in articles in the two papers. You would need to develop criteria or instructions for identifying or measuring criticism. The reliability of your measuring scheme could be assessed by having two people read all the articles, independently rate the level of criticism in them according to your instructions, and then compare their results. Reliability would be demonstrated if both people reached similar conclusions regarding the content of the articles in question.

The reliability of political science measures can be calculated in many different ways. We describe three methods here that are often associated with written test items or survey questions, but the ideas may be applied in other research contexts.

[15] Edward G. Carmines and Richard A. Zeller, *Reliability and Validity Assessment,* A Sage University Paper: Quantitative Applications in the Social Sciences no. 07–017 (Beverly Hills, Calif.: Sage, 1979).

The **test-retest method** involves applying the same "test" to the same observations after a period of time and then comparing the results of the different measurements. For example, if a series of questions measuring liberalism is asked of a group of respondents on two different days, a comparison of their scores at both times could be used as an indication of the reliability of the measure of liberalism. We frequently engage in test-retest behavior in our everyday lives. How often have you stepped on the bathroom scale twice in a matter of seconds?

The test-retest method of measuring reliability may be both difficult and problematic, since one must measure the phenomenon at two different points. It is possible that two different results may be obtained because what is being measured has changed, not because the measure is unreliable. For example, if your bathroom scale gives you two different weights within a few seconds, the scale is unreliable, as your weight cannot have changed. However, if you weigh yourself once a week for a month and find that you get different results each time, is the scale unreliable, or has your weight changed between measurements? A further problem with the test-retest check for reliability is that the administration of the first measure may affect the second measure's results. For instance, the difference between SAT Reasoning Test scores the first and second times that individuals take the test may not be assumed to be a measure of the reliability of the test, since test takers might alter their behavior the second time as a result of taking the test the first time (e.g., they might learn from their first experience with the test).

The **alternative-form method** of measuring reliability also involves measuring the same attribute more than once, but it uses two different measures of the same concept rather than the same measure. For example, a researcher could devise two different sets of questions to measure the concept of liberalism, ask the same respondents questions at two different times using one set of questions the first time and the other set of questions the second time, and compare the respondents' scores. Using two different forms of the measure reduces the chance that the second scores are influenced by the first measure, but it still requires the phenomenon to be measured twice. Depending on the length of time between the two measurements, what is being measured may change.

The **split-halves method** of measuring reliability involves applying two measures of the same concept at the same time. The results of the two measures are then compared. This method avoids the problem that the concept being measured may change between measures. The split-halves method is often used when a multi-item measure can be split into two equivalent halves. For example, a researcher may devise a measure of liberalism consisting of the responses to ten questions on a public opinion survey. Half of these questions could be selected to represent one measure of liberalism, and the other half selected to represent a second measure of liberalism. If individual scores on the two measures of liberalism are similar, then the ten-item measure may be said to be reliable by the split-halves approach.

The test-retest, alternative-form, and split-halves methods provide a basis for calculating the similarity of results of two or more applications of the same or equivalent measures. The less consistent the results are, the less reliable the measure. Political scientists take very seriously the reliability of the measures they use. Survey researchers are often concerned about the reliability of the answers they receive. For example, respondents' answers to survey questions often vary considerably when the instruments are given at two different times.[16] If respondents are not concentrating or taking the survey seriously, the answers they provide may as well have been pulled out of a hat.

Now, let us return to the example of measuring your weight using a home scale. If you weigh yourself on your home scale, then go to the gym and weigh yourself again there, and get the same number (alternative forms test of reliability), you may conclude that your home scale is reliable. But what if you get two different numbers? Assuming your weight has not changed, what is the problem? If you go back home immediately and step back on your home scale and find that it gives you a measurement that is different from the first it gave you, you could conclude that your scale has a faulty mechanism, is inconsistent, and therefore is unreliable. However, what if your bathroom scale gives you the same weight as the first time? It would appear to be reliable. Maybe the gym scale is unreliable. You could test this out by going back to the gym and reweighing yourself. If the gym scale gives a reading different from the one it gave the first time, then it is unreliable. But what if the gym scale gives consistent readings? Each scale appears to be reliable (the scales are not giving you different weights at random), but at least one of them is giving you a wrong measurement (that is, not giving you your correct weight). This is a problem of validity.

Validity

Essentially, a valid measure is one that measures what it is supposed to measure. Unlike reliability, which depends on whether repeated applications of the same or equivalent measures yield the same result, **validity** refers to the degree of correspondence between the measure and the concept it is thought to measure.

Let us consider first an example of a measure whose validity has been questioned: voter turnout. Many studies examine the factors that affect voter turnout and, thus,

[16] Philip E. Converse, "The Nature of Belief Systems in Mass Publics," in *Ideology and Discontent,* ed. David E. Apter (New York: Free Press of Glencoe, 1964); Pauline Marie Vaillancourt, "Stability of Children's Survey Responses," *Public Opinion Quarterly* 37, no. 3 (1973): 373–87; J. Miller McPherson, Susan Welch, and Cal Clark, "The Stability and Reliability of Political Efficacy: Using Path Analysis to Test Alternative Models," *American Political Science Review* 71, no. 2 (1977): 509–21; and Philip E. Converse and Gregory B. Markus, "Plus ça change . . . : The New CPS Election Study Panel," *American Political Science Review* 73, no. 1 (1979): 32–49.

require an accurate measurement of voter turnout. One way of measuring voter turnout is to ask people if they voted in the last election—self-reported voting. However, given the social desirability of voting in the United States—wearing the "I voted" sticker or posting "I voted" on a social media site can bring social rewards—will nonvoters admit their failure to vote to an interviewer? Some nonvoters may claim in surveys to have voted, resulting in an invalid measure of voter turnout that overstates the number of voters. In fact, this is what usually happens. Voter surveys commonly overestimate turnout by several percentage points.[17]

A measure can also be invalid if it measures a slightly or very different concept than intended. For example, assume that a researcher intends to measure ideology, conceptualized as an individual's political views on a continuum between conservative, moderate, and liberal. The researcher proposes to measure ideology by asking survey respondents, "To which party do you feel closest, the Democratic Party or the Republican Party?" This measure would be invalid because it fails to measure ideology as conceptualized. Partisan affinity, while often consistent with ideology, is not the same as ideology. This measure could be a valid measure of party identification, but not ideology.

A measure's validity is more difficult to demonstrate empirically than is its reliability because validity involves the relationship between the measurement of a concept and the actual presence or amount of the concept itself. Information regarding the correspondence is seldom abundant. Nonetheless, there are ways to evaluate the validity of any particular measure. In the following paragraphs we explain several ways of thinking about validity including face, content, construct, and interitem validity.

Face validity may be asserted (not empirically demonstrated) when the measurement instrument appears to measure the concept it is supposed to measure. To assess the face validity of a measure, we need to know the meaning of the concept being measured and whether the information being collected is "germane to that concept."[18] For example, let us return to thinking about how we might measure political ideology—that is, whether someone is conservative, moderate, or liberal. Such a measure could be as simple as a question used by the Pew Research Center: "Do you think of yourself as conservative, moderate or liberal?"[19] On its face, this measure appears to capture the intended concept, so it has face validity. It might be tempting to use individuals' responses to a question on party identification, but one would be assuming that all Democrats are liberal and all Republicans

[17] Raymond E. Wolfinger and Steven J. Rosenstone, appendix A in *Who Votes?* (New Haven, Conn.: Yale University Press, 1980).

[18] Kenneth D. Bailey, *Methods of Social Research* (New York: Free Press, 1978), 58.

[19] Pew Research Center. Retrieved December 5, 2014. Available at http://www.pewresearch.org/data-trend/political-attitudes/political-ideology/

are conservative. Also, if the party identification variable included a category for independents, what would be their ideology? Can you assume they are all moderates? For these reasons, a question measuring party identification would lack face validity as a measure of ideology.

In general, measures lack face validity when there are good reasons to question the correspondence of the measure to the concept in question. In other words, assessing face validity is essentially a matter of judgment. If no consensus exists about the meaning of the concept to be measured, the face validity of the measure is bound to be problematic.

Content validity is similar to face validity but involves determining the full domain or meaning of a particular concept and then making sure that all components of the meaning are included in the measure. For example, suppose you wanted to design a measure of the extent to which a nation's political system is democratic. As noted earlier, *democracy* means many things to many people. Raymond D. Gastil constructed a measure of democracy that included two dimensions, political rights and civil liberties. His checklists for each dimension consisted of eleven items.[20] Political scientists are often interested in concepts with multiple dimensions or complex domains, like democracy, and spend quite a bit of time discussing and justifying the content of their measures. In order for a measure of Gastil's conception of democracy to achieve content validity, the measure should capture all eleven components in the definition.

A third way to evaluate the validity of a measure is by empirically demonstrating **construct validity**. Construct validity can be understood in two different ways: convergent construct validity and divergent construct validity. **Convergent construct validity** is when a measure of a concept is related to a measure of another concept with which the original concept is thought to be related. In other words, a researcher may specify, on theoretical grounds, that two concepts ought to be related in a positive manner (say, political efficacy with political participation or education with income) or a negative manner (say, democracy and human rights abuses). The researcher then develops a measure of each of the concepts and examines the relationship between them. If the measures are positively or negatively correlated, then one measure has convergent validity for the other measure. In the case that there is no relationship between the measures, then the theoretical relationship is in error, at least one of the measures is not an accurate representation of the concept, or the procedure used to test the relationship is faulty. The absence of a hypothesized relationship does not mean a measure is invalid, but the presence of a relationship gives some assurance of the measure's validity.

[20] As discussed in Ross E. Burkhart and Michael S. Lewis-Beck, "Comparative Democracy: The Economic Development Thesis," *American Political Science Review* 88, no. 4 (1994): appendix A.

Discriminant construct validity involves two measures that theoretically are expected *not* to be related; thus, the correlation between them is expected to be low or weak. If the measures do not correlate with one another, then discriminate construct validity is demonstrated.

Let us return to the question of measuring the power of legislative leaders because it provides a good example of the importance of construct validity. As we pointed out before, the perceived-influence approach to measuring power is more difficult to use than the formal-powers approach. Therefore, if the two measures are shown to have construct validity, operationalizing leadership power using the formal-powers approach by itself might be a valid way to measure the concept. If the two measures do not have construct validity, then it would be clear that the two approaches are not measuring the same thing. Thus, which measure is used could greatly affect the findings of research into the factors associated with the presence of strong leadership power or on the consequences of such power. These were the very questions raised by political scientist James Coleman Battista.[21] He constructed several measures of perceived leadership power and correlated them with a measure of formal power. The results, shown in table 5-1, show that the measure of formal power correlates only weakly with three measures of perceived power (which, as expected, correlate well with one another). Therefore, measures of perceived power and the measure of formal power do not demonstrate convergent construct validity.

A fourth way to demonstrate validity is through **interitem association**. This is the type of validity test most often used by political scientists. It relies on the similarity of outcomes of more than one measure of a concept to demonstrate the validity of the entire measurement scheme. It is often preferable to use more than one item to measure a concept—reliance on just one measure is more prone to error or misclassification of a case.[22]

Let us return to the researcher who wants to develop a valid measure of liberalism. First, the researcher might measure people's attitudes toward (1) welfare, (2) military spending, (3) abortion, (4) Social Security benefit levels, (5) affirmative action, (6) a progressive income tax, (7) school vouchers, and (8) protection of the rights of the accused. Then the researcher could determine how the responses to each question relate to the responses to each of the other questions. The validity of the measurement scheme would be demonstrated if strong relationships existed among people's responses across the eight questions.

[21] James Coleman Battista, "Formal and Perceived Leadership Power in U.S. State Legislatures," *State Politics and Policy Quarterly* 11, no. 1 (2011).

[22] Joseph A. Gliem and Rosemary R. Gliem, "Calculating, Interpreting, and Reporting Cronbach's Alpha Reliability Coefficient for Likert-Type Scales" (paper presented at the 2003 Midwest Research to Practice Conference in Adult, Continuing, and Community Education, Ohio State University, Columbus, 2003). Available at https://scholarworks.iupui.edu/bitstream/handle/1805/344/Gliem+&+Gliem.pdf?sequence=1

TABLE 5-1 **Correlations of Leadership Power Measures**

	Formal power index	Average perceived power	% rating highest (all)	% rating highest (internal)
Formal power index	1			
Average perceived power	.186	1		
Prop. rating highest (all)	.086	.698*	1	
Prop. rating highest (internal)	.119	.702*	.684*	1
Combined power index	.915*	.558*	.338*	.375*

*p <.05

Source: James Coleman Battista, "Formal and Perceived Leadership Power in U.S. State Legislatures," *State Politics and Policy Quarterly* 11, no. 1 (2011): tab. 2, p. 209. Copyright © The Author 2011. Reprinted by Permission of SAGE Publications.

Note: We show here only the results for Battista's analysis when southern states are excluded. Because southern states have a history of little two-party competition but may have to contend with party factions, and the question asked respondents to rate the power of legislative leaders as *party* leaders, respondents were likely to give low perceived power ratings. Southern legislative leaders, however, may be strong *chamber* leaders, which is what is being measured in states with a history of two-party competition. See chapter 13 for an explanation of correlation.

The results of such interitem association tests are often displayed in a **correlation matrix**. Such a display shows how strongly related each of the items in the measurement scheme is to all the other items. In the hypothetical data shown in table 5-2, we can see that people's responses to six of the eight measures were strongly related to each other, whereas responses to the questions on protection of the rights of the accused and school vouchers were not part of the general pattern. Thus, the researcher would probably conclude that the first six items all measure liberalism and that, taken together, they are a valid measurement of liberalism.

The figures in table 5-2 are product-moment correlations: numbers that can vary in value from −1.0 to +1.0 and that indicate the extent to which one variable is related to another. The closer the correlation is to ±1, the stronger the relationship; the closer the correlation is to 0.0, the weaker the relationship (see chapter 13 for a full explanation). The figures in the last two rows are considerably closer to 0.0 than are the other entries, indicating that people's answers to the questions about school vouchers and rights of the accused did not follow the same pattern as their answers to the other questions. Therefore, it looks like school vouchers and rights of the accused are not connected to the same concept of liberalism as measured by the other questions.

TABLE 5-2 **Interitem Association Validity Test of a Measure of Liberalism**

	Welfare	Military spending	Abortion	Social Security	Affirmative action	Income tax	School vouchers	Rights of accused
Welfare	1							
Military spending	.56	1						
Abortion	.71	.60	1					
Social Security	.80	.51	.83	1				
Affirmative action	.63	.38	.59	.69	1			
Income tax	.48	.67	.75	.39	.51	1		
School vouchers	.28	.08	.19	.03	.30	−.07	1	
Rights of accused	−.01	.14	−.12	.10	.23	.18	.45	1

Note: Hypothetical data.

Content and face validity are difficult to assess when agreement is lacking on the meaning of a concept, and construct validity, which requires a well-developed theoretical perspective, usually yields a less-than-definitive result. The interitem association test requires multiple measures of the same concept. Although these validity "tests" provide important evidence, none of them is likely to support an unequivocal decision concerning the validity of particular measures.

Problems with Reliability and Validity in Political Science Measurement

Survey researchers often want to measure respondents' household income. Measurement of this basic variable illustrates the numerous threats to the reliability and validity of political science measures. The following is a question used in the 2004 American National Election Study (ANES):

> Please look at the booklet and tell me the letter of the income group that includes the income of all members of your family living here in 2003 before taxes. This figure should include salaries, wages, pensions, dividends, interest, and all other income. Please tell me the letter of the income group that includes the income you had in 2003 before taxes.

Respondents were given the following choices:

A. None, or less than $2,999
B. $3,000–$4,999
C. $5,000–$6,999
D. $7,000–$8,999
E. $9,000–$10,999
F $11,000–$12,999
G. $13,000–$14,999
H. $15,000–$16,999
I. $17,000–$19,999
J. $20,000–$21,999
K. $22,000–$24,999
L. $25,000–$29,999

M. $30,000–$34,999
N. $35,000–$39,999
O. $40,000–$44,999
P $45,000–$49,999
Q. $50,000–$59,999
R. $60,000–$69,999
S. $70,000–$79,999
T. $80,000–$89,999
U. $90,000–$104,999
V $105,000—$119,999
W $120,000 and over

Both the reliability and the validity of this method of measuring income are questionable. Threats to the reliability of the measure include the following:

- Respondents may not know how much money they make and therefore incorrectly guess their income.
- Respondents may not know how much money other family members make and guess incorrectly.
- Respondents may know how much they make but carelessly select the wrong categories.
- Interviewers may circle the wrong categories when listening to the selections of the respondents.
- Data-entry personnel may touch the wrong numbers when entering the answers into the computer.
- Dishonest interviewers may incorrectly guess the income of a respondent who does not complete the interview.
- Respondents may not know which family members to include in the income total; some respondents may include only a few family members, while others may include even distant relations.
- Respondents whose income is on the border between two categories may not know which one to pick. Some may pick the higher category; others, the lower one.

Because of these measurement problems, if this measure were applied to the same people at two different times, we could expect the results to vary, resulting in inaccurate measures that are too high for some respondents and too low for others. Some amount of **random measurement error** is likely to occur with any measurement scheme.

In addition to these threats to reliability, there are numerous threats to the validity of this measure:

- Respondents may have illegal income they do not want to reveal and therefore may systematically underestimate their income.
- Respondents may try to impress the interviewer, or themselves, by systematically overestimating their income.
- Respondents may systematically underestimate their before-tax income because they think of their take-home pay and underestimate how much money is being withheld from their paychecks.
- Respondents may systematically skip the question due to privacy concerns over providing a precise number even if they know it.

Notice that this second list of problems contains the word *systematically*. These problems are not simply caused by random inconsistencies in measurements, with some being too high and others too low for unpredictable reasons. *Systematic* measurement error introduces error that may **bias** research results, thus compromising the confidence we have in them.

This long list of problems with both the reliability and the validity of this fairly straightforward measure of a relatively concrete concept is worrisome. Imagine how much more difficult it is to develop reliable and valid measures when the concept is abstract (for example, tolerance, environmental conscience, self-esteem, or liberalism) and the measurement scheme is more complicated.

The reliability and validity of the measures used by political scientists are seldom demonstrated to everyone's satisfaction. Most measures of political phenomena are neither completely invalid or valid nor thoroughly unreliable or reliable but, rather, are partly accurate. Therefore, researchers generally present the rationale and evidence available in support of their measures and attempt to persuade their audience that their measures are at least as accurate as alternative measures would be. Nonetheless, a skeptical stance on the part of the reader toward the reliability and validity of political science measures is often warranted.

Note, finally, that reliability and validity are not the same thing. A measure may be reliable without being valid. One may devise a series of questions to measure liberalism, for example, that yields the same result for the same people every time but that misidentifies individuals. A valid measure, however, will also be reliable: if it accurately measures the concept in question, then it will do so consistently across measurements—allowing, of course, for some random measurement error that may occur. It is more important, then, to demonstrate validity than reliability, but reliability is usually more easily and precisely tested.

The Precision of Measurements

Measurements should be not only accurate but also precise; that is, measurements should contain as much information as possible about the attribute or behavior being measured. The more precise our measures, the more complete and informative can be our test of the relationships between two or more variables.

Suppose, for example, that we wanted to measure the height of political candidates to see if taller candidates usually win elections. Height could be measured in many different ways. We could have two categories of the variable "height"—*tall* and *short*—and assign different candidates to the two categories based on whether they were of above-average or below-average height. Or we could compare the heights of candidates running for the same office and measure which candidate was the tallest, which the next tallest, and so on. Or we could take a tape measure and measure each candidate's height in inches and record that measure. The last method of measurement captures the most information about each candidate's height and is, therefore, the most precise measure of the attribute.

Levels of Measurement

When we consider the precision of our measurements, we refer to the **level of measurement**. The level of measurement involves the type of information that we think our measurements contain and the mathematical properties that determine the type of comparisons that can be made across a number of observations on the same variable. The level of measurement also refers to the claim we are willing to make when we assign numbers to our measurements.

There are four different levels of measurement: nominal, ordinal, interval, and ratio. While few concepts used in political science research inherently require a particular level of measurement, there are methodological limitations because some measures provide more information and better mathematical properties than others. So the level of measurement used to measure any particular concept is a function of both the researcher's imagination and resources, and methodological needs.

We begin with **nominal measurement**, the level that has the fewest mathematical properties of the four levels. A nominal-level measure indicates that the values assigned to a variable represent only different categories or classifications for that variable. In such a case, no category is more or less than another category; they are simply different. For example, suppose we measure the religion of individuals by asking them to indicate whether they are Christian, Jewish, Muslim, or other. Since the four categories or values for the variable religion are simply different, the measurement is at a nominal level. Other common examples of nominal-level measures are gender, marital status, and state of residence. A nominal measure of partisan affiliation might have the following categories: Democrat, Republican, Green, Libertarian, other, and none. Numbers will be assigned to the categories when the data are coded for statistical analysis, but these numbers do not represent mathematical differences between the categories—any of the parties could be assigned any number, as long as those numbers are different from each other. In this sense, nominal-level measures provide the least amount of information about a concept. An **ordinal measurement** has all of the properties of a nominal measure but also assumes observations can be compared in terms of having more or less of a particular attribute. Hence, the ordinal level of measurement captures more information about the measured concept and has more mathematical properties than a nominal-level measure. For example, we could create an ordinal measure of formal education completed with the following categories: "eighth grade or less," "some high school," "high school graduate," "some college," and "college degree or more." Here we are concerned not with the exact difference between the categories of education but only with whether one category is more or less than another. When coding this variable, we would assign higher numbers to higher categories of education. The intervals between the numbers have no meaning; all that matters is that the higher numbers represent more of the attribute than do the lower numbers. An ordinal variable measuring partisan affiliation with the categories "strong Republican," "weak Republican," "neither leaning Republican nor Democrat," "weak Democrat," and "strong Democrat" could be assigned codes 1, 2, 3, 4, 5 or 1, 2, 5, 8, 9 or any other combination of numbers, as long as they were in ascending or descending order.

Dichotomous nominal variables—that is, nominal-level variables with only two categories—are nominal-level measures, but frequently treated as ordinal-level measures. For example, we could measure nuclear capability with two categories, where a country that has nuclear capabilities would be coded as a one and a country that does not would be coded as a zero. One could interpret this variable as nuclear capability being present or absent in a country and therefore a one represents more of the concept, nuclear capability. To give another example, a person who did not vote in the last election lacks, or has less of, the attribute of having voted than a person who did vote.

Because nominal and ordinal measures rely on categories it is important to make sure these variables are both exhaustive and exclusive. *Exhaustive* refers to making sure that all possible categories—or answer choices—are accounted for. The simplest solution to make sure a variable is exhaustive is to include an "other" category that can be used for values that are not represented in the identified categories. *Exclusive* refers to making sure that a single value or answer can only fit into one category. Each category should be distinct from the others, with no overlap.

HELPFUL HINTS

Debating the Level of Measurement

Suppose we ask individuals three questions designed to measure social trust, and we believe that individuals who answer all three questions a certain way have more social trust than persons who answer two of the questions a certain way, and these individuals have more social trust than individuals who answer one of the questions a certain way. We could assign a score of 3 to the first group, 2 to the second group, 1 to the third group, and 0 to those who did not answer any of the questions in a socially trusting manner. In this case, the higher the number, the more social trust an individual has.

What level of measurement is this variable? It might be considered to be ratio level, if one interprets the variable as simply the number of questions answered indicating social trust. But does a person who has a score of 0 have no social trust? Does a person with a score of 3 have three times as much social trust as a person with a score of 1? Perhaps, then, the variable is an interval-level measure, if one is willing to assume that the difference in social trust between individuals with scores of 2 and 3 is the same as the difference between individuals with scores of 1 and 2. But what if the effect of answering more questions in the affirmative is not simply additive? In other words, perhaps a person who has a score of 3 has a lot more social trust than someone with a score of 2 and that this difference is more than the difference between individuals with scores of 1 and 2. In this case, then, the measure would be ordinal level, not interval level.

Check out more Helpful Hints at **edge.sagepub.com/johnson8e**

The next level of measurement, an **interval measurement**, includes the properties of the nominal level (characteristics are different) and the ordinal level (characteristics can be put in a meaningful order). But unlike the preceding levels of measurement, the intervals between the categories or values assigned to the observations do have meaning. The value of a particular observation is important not just in terms of whether it is larger or smaller than another value (as in ordinal measures) but also in terms of *how much* larger or smaller it is. For example, suppose we record the year in which certain events occurred. If we have three observations—1950, 1962, and 1977—we know that the event in 1950 occurred twelve years before the one in 1962 and twenty-seven years before the one in 1977. A one-unit change (the interval) all along this measurement is identical in meaning: the passage of one year's time.

Another characteristic of an interval level of measurement that distinguishes it from the next level of measurement (ratio) is that an interval-level measure has an arbitrarily assigned zero point that does not represent the absence of the attribute being measured. For example, many time and temperature scales have arbitrary zero points. Thus, the year 0 CE does not indicate the beginning of time—if this were true, there would be no BCE dates. Nor does 0°C indicate the absence of heat; rather, it indicates the temperature at which water freezes. For this reason, with interval-level measurements we cannot calculate ratios; that is, we cannot say that 60°F is twice as warm as 30°F. So while the interval level of measurement captures more information and mathematical properties than the nominal and ordinal levels, it does not have the full properties of mathematics.

The final level of measurement is a **ratio measurement**. This type of measurement involves the full mathematical properties of numbers and contains the most possible information about a measured concept. That is, a ratio-level measure includes the values of the categories, the order of the categories, and the intervals between the categories; it also precisely indicates the relative amounts of the variable that the categories represent because its scale includes a meaningful zero. If, for example, a researcher is willing to claim that an observation with ten units of a variable possesses exactly twice as much of that attribute as an observation with five units of that variable, then a ratio-level measurement exists. The key to making this assumption is that a value of zero on the variable actually represents the absence of that variable. Because ratio measures have a true zero point, it makes sense to say that one measurement is *x* times another. It makes sense to say a sixty-year-old person is twice the age of a thirty-year-old person (60/30 = 2), whereas it does not make sense to say that 60°C is twice as warm as 30°C.[23]

Political science researchers have measured many concepts at the ratio level. People's ages, unemployment rates, percentage of the vote for a particular candidate, and crime rates are all measures that contain a zero point and possess the full mathematical properties of the numbers used. However, more political science research has probably relied on nominal- and ordinal-level measures than on interval- or ratio-level measures. This has restricted the types of hypotheses and analysis techniques that political scientists have been willing and able to use.

Identifying the level of measurement of variables is important, since it affects the data analysis techniques that can be used and the conclusions that can be drawn about the relationships between variables. Higher-order methods often require higher levels of measurement, while other methods have been developed for lower levels

[23] The distinction between an interval-level and a ratio-level measure is not always clear, and some political science texts do not distinguish between them. Interval-level measures in political science are rather rare; ratio-level measures (money spent, age, number of children, years living in the same location, for example) are more common.

of measurement. The decision of which level of measurement to use is not always a straightforward one, and uncertainty and disagreement often exist among researchers concerning these decisions. Few phenomena inherently require one particular level of measurement. Often, a phenomenon can be measured with any level of measurement, depending on the particular technique designed by the researcher and the claims the researcher is willing to make about the resulting measure.

Working with Precision: Too Little or Too Much

Researchers usually try to devise as high a level of measurement for their concepts as possible (nominal being the lowest level of measurement and ratio the highest). With a higher level of measurement, more advanced data analysis techniques can be used, and more precise statements can be made about the relationships between variables. Thus, researchers measuring attitudes or concepts with multiple operational definitions often construct a scale or an index from nominal-level measures that permits at least ordinal-level comparisons between observations. We discuss the construction of indexes and scales in greater detail in the following paragraphs.

It is easy to transform ratio-level information (e.g., age in number of years) into ordinal-level information (e.g., age groups). However, if you start with the ordinal-level measure, age groups, you will not have each person's actual age. If you decide you want to use a person's actual age, you will have to collect that data—it cannot be created from an ordinal-level measurement. Similarly, a researcher investigating the effect of campaign spending on election outcomes could use a ratio-level variable measuring how much each candidate spent on his or her campaign. This information could be used to construct a new variable indicating how much more one candidate spent than the other, or simply whether or not a candidate spent more than his or her opponent. Candidate spending could also be grouped into ranges.

Nominal and ordinal variables with many categories or interval- and ratio-level measures using more decimal places are more precise than measures with fewer categories or decimal places, but sometimes the result may provide more information than can be used. Researchers frequently start out with ratio-level measures or with ordinal and nominal measures with quite a few categories but then collapse or combine the data to create groups or fewer categories. They do this so that they have enough cases in each category for statistical analysis or to make comparisons easier to follow. For example, one might want to present comparisons simply between Democrats and Republicans rather than presenting data broken down into categories of strong, moderate, and weak for each party.

It may seem contradictory now to point out that extremely precise measures also may create problems. For example, measures with many response possibilities take up space if they are questions on a written questionnaire or require more time to explain if they are included in a telephone survey. Such questions may also confuse or tire survey respondents. A more serious problem is that they may lead to measurement error. Think about the possible responses to a question asking respondents to use a 100-point scale (called a thermometer scale) to indicate their support for or opposition to a political candidate, assuming that 50 is considered the neutral position and 0 is least favorable or "coldest" and 100 most favorable. Some respondents may not use the whole scale (to them, no candidate ever deserves more than an 80 or less than a 20), whereas others may use the ends and the very middle of the scale and ignore the scores in between. We might predict that a person who gives a candidate a 100 is more likely to vote for that candidate than is a person who gives the same candidate an 80, but in reality they may like the candidate pretty much the same way and would be equally likely to vote for the candidate. Another problem with overly precise measurements is that they may be unreliable. If asked to rate candidates on more than one occasion, respondents could vary slightly the number that they choose, even if their opinion has not changed.

Multi-Item Measures

Many measures consist of a single item. For example, the measures of party identification, whether or not one party controls Congress, the percentage of the vote received by a candidate, how concerned about an issue a person is, the policy area of a judicial case, and age are all based on a single measure of each phenomenon in question. Often, however, researchers need to devise measures of more complicated phenomena that have more than one facet or dimension. For example, internationalism, political ideology, political knowledge, dispersion of political power, and the extent to which a person is politically active are complex phenomena or concepts that may be measured in many different ways.

In this situation, researchers often develop a measurement strategy that allows them to capture numerous aspects of a complex phenomenon while representing the existence of that phenomenon in particular cases with a single representative value. Usually this involves the construction of a multi-item index or scale representing the several dimensions of the phenomenon. These multi-item measures are useful because they enhance the accuracy of a measure, simplify a researcher's data by reducing them to a more manageable size, and increase the level of measurement of a phenomenon. In the remainder of this section, we describe several common types of indexes and scales.

Indexes

A **summation index** is a method of accumulating scores on individual items to form a composite measure of a complex phenomenon. An index is constructed by assigning a range of possible scores for a certain number of items, determining the score for each item for each observation, and then combining the scores for each observation across all the items. The resulting summary score is the representative measurement of the phenomenon.

A researcher interested in measuring how much freedom exists in different countries, for example, might construct an index of political freedom by devising a list of items germane to the concept, determining where individual countries score on each item, and then adding these scores to get a summary measure. In table 5-3, such a hypothetical index is used to measure the amount of freedom in countries A through E.

The index in table 5-3 is a simple, additive one; that is, each item counts equally toward the calculation of the index score, and the total score is the summation of the individual item scores. However, indexes may be constructed with more complicated aggregation procedures and by counting some items as more important than others. In the preceding example, a researcher might consider some indicators of freedom as more important than others and wish to have them contribute more to the calculation of the final index score. This could be done either by weighting

TABLE 5-3 Hypothetical Index for Measuring Freedom in Countries

Does the country possess:	Country A	Country B	Country C	Country D	Country E
Privately owned newspapers	1	0	0	0	1
Legal right to form political parties	1	1	0	0	0
Contested elections for significant public offices	1	1	0	0	0
Voting rights extended to most of the adult population	1	1	0	1	0
Limitations on government's ability to incarcerate citizens	1	0	0	0	1
Index score	5	3	0	1	2

Note: Hypothetical data. The score is 1 if the answer is yes, 0 if no.

(multiplying) some item scores by a number indicating their importance or by assigning a higher score than 1 to those attributes considered more important.

Indexes are often used with public opinion surveys to measure political attitudes. This is because attitudes are complex phenomena and we usually do not know enough about them to devise single-item measures. So we often ask several questions of people about a single attitude and aggregate the answers to represent the attitude. A researcher might measure attitudes toward abortion, for example, by asking respondents to choose one of five possible responses—strongly agree, agree, undecided, disagree, and strongly disagree—to the following three statements: (1) Abortions should be permitted in the first three months of pregnancy. (2) Abortions should be permitted if the woman's life is in danger. (3) Abortions should be permitted whenever a woman wants one.

An index of attitudes toward abortion could be computed by assigning numerical values to each response (such as 1 for *strongly agree*, 2 for *agree*, 3 for *undecided*, and so on) and then adding the values of a respondent's answers to these three questions. (The researcher would have to decide what to do when a respondent did not answer one or more of the questions.) The lowest possible score in this case would be a 3, indicating the most extreme pro-abortion attitude, and the highest possible score would be a 15, indicating the most extreme anti-abortion attitude. Scores in between would indicate varying degrees of approval of abortion.

Indexes are typically fairly simple ways of producing single representative scores of complicated phenomena such as political attitudes. They are probably more accurate than most single-item measures, but they may also be flawed in important ways. Aggregating scores across several items assumes, for example, that each item is equally important to the summary measure of the concept and that the items used faithfully encompass the domain of the concept. Although individual item scores can be weighted to change their contribution to the summary measure, the researcher often has little information upon which to base a weighting scheme.

Several standard indexes are often used in political science research. The FBI crime index, the Consumer Confidence Index, and the Consumer Price Index, for example, have been used by many researchers. Before using these or any other readily available index, you should familiarize yourself with its construction and be aware of any questions raised about its validity. Although simple summation indexes are generally more accurate than single-item measures of complicated phenomena, it is often unclear how valid they are or what level of measurement they represent. For example, is the index of freedom an ordinal-level measure, or could it be an interval-level or even a ratio-level measure? Another possible issue with indexes such as the Consumer Price Index is that what goes into its calculation can change over time.[24]

[24] Brett Arends, "Why You Can't Trust the Inflation Numbers," *Wall Street Journal*, January 26, 2011, http://online.wsj.com/article/SB10001424052748704013604576104351050317610.html

HELPFUL HINTS

Creating an Index of Speakers' Institutional Powers

Richard A. Clucas created an index of the institutional power of state house speakers by sorting powers into five general categories and assigning values to powers within those categories using the following scoring procedure:

- Appointment power index

  Speaker selects all chairs and party leaders = 5 points

  Selects chairs and a majority of leaders = 4

  Selects chairs; selects few or no leaders = 3

  Shares powers to select chairs; selects few or no leaders = 2

  Does not select chairs; selects few or no leaders = 1

- Committee Power Index

  Speaker assigns all members to committee; decides number of committees = 5.0

  Assigns all members; shares power over number = 4.5

  Assigns all members; does not decide number = 4.0

  Assigns majority members; decides number = 3.5

  Assigns majority members; shares power over number = 3.0

  Assigns majority members; does not decide number = 2.5

  Another actor has formal power over assignments, but the speaker shares in process, such as serving as a member of rules committee; decides committee number = 2.0

  Shares in assignment process; shares power over number = 1.5

  Shares in assignment process; does not decide number = 1.0

  Not involved in either = 0.0

- Resource Power Index

  Campaign committee exists; speaker has control over legislative employees = 5.0

  Committee exists; speaker does not control employees = 3.0

  Committee does not exist; speaker controls employees = 1.0

  Committee does not exist; speaker does not control employees = 0.0

- Procedural Power Index: The index was created by first creating separate indexes for the speaker's power over bill referral and floor procedures. The average of these scores was then used for the index.

- Bill Referral Power Index

  Speaker has complete control over bill referral = 5

  Controls referral, but there are restrictions on its use = 4

  Shares power over referral; no restrictions on use = 3

  Shares power over referral; actions restricted = 2

  Not involved in referral = 1

- Floor Powers Index

  Speaker prepares calendar, decides question, directs chamber = 5

  Controls two of these floor powers = 4

  Controls one of these floor powers = 3

  Has no control over floor = 1

- Tenure Power Index

  No tenure limit = 5

  Eight years = 4

  Six years = 3

  Four years = 2

  Two years = 1

Clucas's overall index can range in value from 3 to 25. Each category has a maximum value of 5; thus, he is making the claim that full power in each category is of equal significance. He uses the index as if it were an interval-level measure, which means that a unit change in any power category is equivalent. These assumptions are reasonable but are certainly open to challenge.

Source: Richard A. Clucas, "Principle-Agent Theory and the Power of State House Speakers," *Legislative Studies Quarterly* 26, no. 2 (2001): 319–38; material taken from the appendix on pp. 335 and 336.

Scales

Although indexes are generally an improvement over single-item measures, their construction also contains an element of arbitrariness. Both the selection of particular items making up the index and the way in which the scores on individual items are aggregated are based on the researcher's judgment. Scales are also multi-item measures, but the selection and combination of items in them is accomplished more systematically than is usually the case for indexes. Over the years, several different kinds of multi-item scales have been used frequently in political science research. We discuss three of them: Likert scales, Guttman scales, and Mokken scales.

A **Likert scale** score is calculated from the scores obtained on individual items. Each item generally asks a respondent to indicate a degree of agreement or disagreement with the item, as with the abortion questions discussed earlier. A Likert scale differs from an index, however, in that once the scores on each of the items are obtained, only some of the items are selected for inclusion in the calculation of the final score. Those items that allow a researcher to distinguish most readily those scoring high on an attribute from those scoring low will be retained, and a new scale score will be calculated based only on those items.

For example, consider the researcher interested in measuring the liberalism of a group of respondents. Since definitions of *liberalism* vary, the researcher cannot be sure how many aspects of liberalism need to be measured. With Likert scaling, the researcher would begin with a large group of questions thought to express various aspects of liberalism with which respondents would be asked to agree or disagree. A provisional Likert scale for liberalism, then, might look like the one in table 5-4.

In practice, a set of questions like this would be scattered throughout a questionnaire so that respondents do not see them as related. Some of the questions might also be worded in the opposite way (that is, so an "agree" response is a conservative response) to ensure genuine answers.

The respondents' answers to these eight questions would be summed to produce a provisional score. The scores in this case can range from 8 to 40. Then the

TABLE 5-4 **Provisional Likert Scale to Measure Concept of Liberalism**

	Strongly Disagree (1)	Disagree (2)	Undecided (3)	Agree (4)	Strongly Agree (5)
The government should ensure that no one lives in poverty.	—	—	—	—	—
Military spending should be reduced.	—	—	—	—	—
It is more important to take care of people's needs than it is to balance the federal budget.	—	—	—	—	—
Social Security benefits should not be cut.	—	—	—	—	—
The government should spend money to improve housing and transportation in urban areas.	—	—	—	—	—
Wealthy people should pay taxes at a much higher rate than poor people.	—	—	—	—	—
Busing should be used to integrate public schools.	—	—	—	—	—
The rights of persons accused of a crime must be vigorously protected.	—	—	—	—	—

responses of the most liberal and the most conservative people to each question would be compared; any questions with similar answers from the disparate respondents would be eliminated—such questions would not distinguish liberals from conservatives. A new summary scale score for all the respondents would be calculated from the questions that remained. A statistic called Cronbach's alpha, which measures internal consistency of the items in the scale and has a maximum value of 1.0, is used to determine which items to drop from the scale. The rule of thumb is that Cronbach's alpha should be 0.8 or above; items are dropped from the scale one at a time until this value is reached.[25]

Likert scales are improvements over multi-item indexes because the items that make up the multi-item measure are selected in part based on the respondents' behavior rather than on the researcher's judgment. Likert scales suffer two of the other defects of indexes, however. The researcher cannot be sure that all the dimensions of a concept have been measured, and the relative importance of each item is still determined arbitrarily.

The **Guttman scale** also uses a series of items to produce a scale score for respondents. Unlike the Likert scale, however, a Guttman scale presents respondents with a range of attitude choices that are increasingly difficult to agree with; that is, the items composing the scale range from those easy to agree with to those difficult to agree with. Respondents who agree with one of the "more difficult" attitude items will also generally agree with the "less difficult" ones. (Guttman scales have also been used to measure attributes other than attitudes. Their main application has been in the area of attitude research, however, so an example of that type is used here.)

Let us return to the researcher interested in measuring attitudes toward abortion. He or she might devise a series of items ranging from "easy to agree with" to "difficult to agree with." Such an approach might be represented by the following items:

Do you agree or disagree that abortions should be permitted:

1. When the life of the woman is in danger
2. In the case of incest or rape
3. When the fetus appears to be unhealthy
4. When the father does not want to have a baby
5. When the woman cannot afford to have a baby
6. Whenever the woman wants one

[25] Gliem and Gliem, "Calculating, Interpreting, and Reporting Cronbach's Alpha Reliability Coefficient for Likert-Type Scales."

This array of items seems likely to result in responses consistent with Guttman scaling. A respondent agreeing with any one of the items is likely also to agree with those items numbered lower than that one. This would result in the "stepwise" pattern of responses characteristic of a Guttman scale.

Suppose six respondents answered this series of questions, as shown in table 5-5. Generally speaking, the pattern of responses is as expected; those who agreed with the "most difficult" questions were also likely to agree with the "less difficult" ones. However, the responses of three people (2, 4, and 5) to the question about the father's preferences do not fit the pattern. Consequently, the question about the father does not seem to fit the pattern and would be removed from the scale. Once that has been done, the stepwise pattern becomes clear.

With real data, it is unlikely that every respondent would give answers that fit the pattern perfectly. For example, in table 5-5, respondent 6 gave an "agree" response to the question about incest or rape. This response is unexpected and does not fit the pattern. Therefore, we would be making an error if we assigned a scale score of 0 to respondent 6. When the data fit the scale pattern well (number of errors is small), researchers assume that the scale is an appropriate measure and that the respondent's "error" may be "corrected" (in this case, either the "agree" in the case of incest or rape or the "disagree" in the case of the life of the woman). There are standard procedures to follow to determine how to correct the data to make it conform to the scale pattern. We emphasize, however, that this is done only if the changes are few.

Guttman scales differ from Likert scales in that, in the former case, generally only one set of responses will yield a particular scale score. That is, to get a score of 3

TABLE 5-5 **Guttman Scale of Attitudes toward Abortion**

Respondent	Life of woman	Incest or rape	Unhealthy fetus	Father	Afford	Anytime	No. of agree answers	Revised scale score
1	A	A	A	A	A	A	6	5
2	A	A	A	D	A	D	4	4
3	A	A	A	D	D	D	3	3
4	A	A	D	A	D	D	3	2
5	A	D	D	A	D	D	2	1
6	D	A	D	D	D	D	1	0

Note: Hypothetical data. A = Agree, D = Disagree.

on the abortion scale, a particular pattern of responses (or something very close to it) is necessary. In the case of a Likert scale, however, many different patterns of responses can yield the same scale score. A Guttman scale is also much more difficult to achieve than a Likert scale, since the items must have been ordered and be perceived by the respondents as representing increasingly more difficult responses reflecting the same attitude.

Both Likert and Guttman scales have shortcomings in their level of measurement. The level of measurement produced by Likert scales is, at best, ordinal (since we do not know the relative importance of each item and so cannot be sure that a 5 answer on one item is the same as a 5 answer on another), and the level of measurement produced by Guttman scales is usually assumed to be ordinal.

Another type of scaling procedure, called Mokken scaling, also analyzes responses to multiple items by respondents to see if, for each item, respondents can be ordered and if items can be ordered.[26] The **Mokken scale** was used by Saundra K. Schneider, William G. Jacoby, and Daniel C. Lewis to see if there was structure and coherence in public opinion regarding the distribution of responsibilities between the federal government and state and local governments.[27] Respondents were asked whether they thought state or local governments "should take the lead" rather than the national government for thirteen different policy areas. The scaling procedure allowed the researchers to see if a specific sequence of policies emerged while moving from one end of the scale to the other. One end of the scale would indicate maximal support for national policy activity, while the other end would indicate maximal support for subnational government policy responsibility.

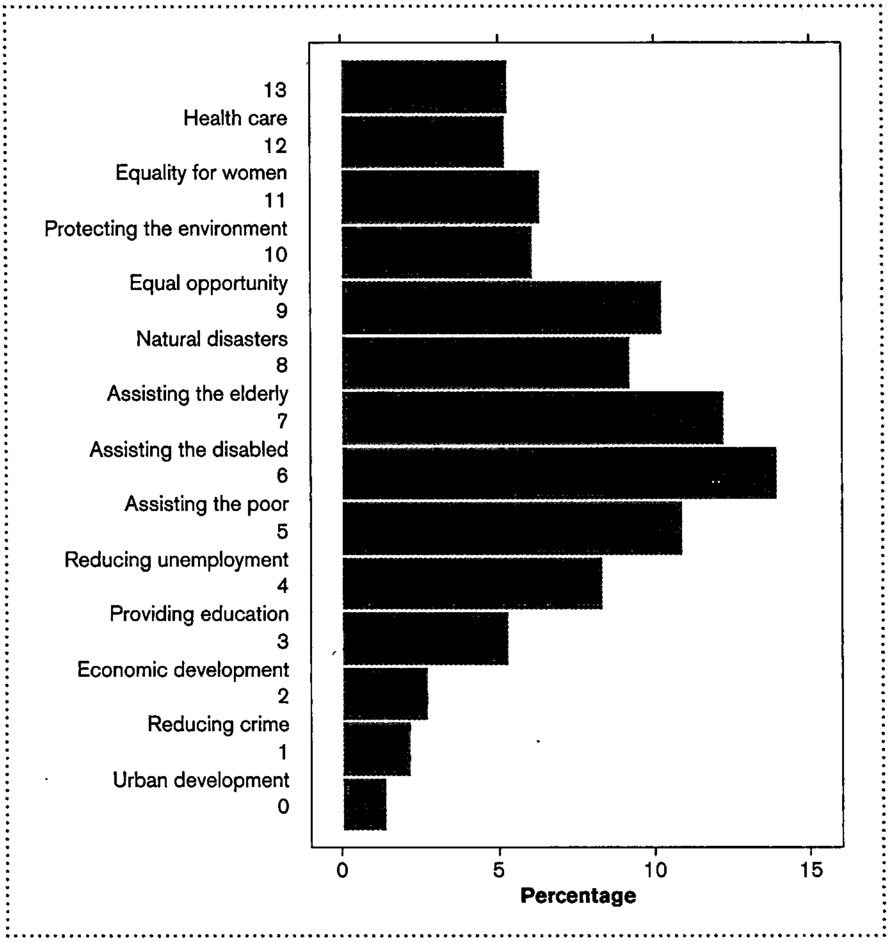
The results of their analysis are shown in figure 5-1. The scale runs from 0 to 13, with 0 indicating that the national government should take the lead in all thirteen policy areas and a score of 13 indicating that the respondent believes that state and local governments should take the lead in all policy areas. A person at any scale score believes that the state and local governments should take the lead in all policy areas that fall below that score. Thus, a person with a score of 9 believes that the national government should take the lead in health care, equality for women, protecting the environment, and equal opportunity, and that state and local governments should take the lead responsibility for natural disasters down to urban development. The bars in the figure correspond to the percentage of respondents

[26] Robert Jan Mokken, *A Theory and Procedure of Scale Analysis with Applications in Political Research* (The Hague, Neth.: Mouton, 1971). Mokken scaling is an example of a nonparametric item response theory (IRT) model. It differs from Guttman scaling in that it has a probabilistic interpretation, whereas Guttman scaling does not. See Ate Dijktra, Girbe Buist, Peter Moorer, and Theo Dassen, "Construct Validity of the Nursing Care Dependency Scale," *Journal of Clinical Nursing* 8, no. 4 (1999): 380–88.

[27] Saundra K. Schneider, William G. Jacoby, and Daniel C. Lewis, "Public Opinion toward Intergovernmental Policy Responsibilities," *Publius: The Journal of Federalism* 41, no. 1 (2011): 1–30.

who received a particular score. Thus, just slightly more than 5 percent of the respondents thought that state and local governments should take the lead in all policy areas, whereas only 1 percent of the respondents thought the national government should take the lead in all thirteen policy areas. Most respondents divided up the responsibilities. The authors concluded from their analysis that the public

FIGURE 5-1 **Mokken Scale of Public Opinion about National versus State/Local Government Responsibilities for Specific Policy Areas**



Source: 2006 Comparative Congressional Election Survey, cited in Saundra K. Schneider, William G. Jacoby, and Daniel C. Lewis, "Public Opinion toward Intergovernmental Policy Responsibilities," *Publius: The Journal of Federalism* 41, no. 1 (2011): fig. 3, p. 14. Reprinted by permission of Oxford University Press. © The Author 2010. Published by Oxford University Press on behalf of CSF Associates: Publius, Inc.

does have a rationale behind its preferences for the distribution of policy responsibilities between national versus state and local governments.

The procedures described so far for constructing multi-item measures are fairly straightforward. There are other advanced statistical techniques for summarizing or combining individual items or variables. For example, it is possible that several variables are related to some underlying concept. **Factor analysis** is a statistical technique that may be used to uncover patterns across measures. It is especially useful when a researcher has a large number of measures and when there is uncertainty about how the measures are interrelated.

An example is the analysis by Daniel D. Dutcher, who conducted research on the attitudes of owners of streamside property toward the water-quality improvement strategy of planting trees in a wide band (called riparian buffers) along the sides

TABLE 5-6 Items Measuring Landowner Attitudes toward Riparian Buffers on Their Streamside Property, Sorted into Dimensions Using Factor Analysis

Maintaining Property Aesthetics

- Maintaining my view of the stream
- Maintaining the look of a pastoral or meadow stream
- Maintaining open space
- Wanting the neighbors to think that I'm doing my part to keep up the traditional look of the neighborhood

Contributing to Stream and Bay Quality

- Being confident that maintaining or creating a streamside forest on my property is necessary to protect the stream
- Contributing to the improvement of downstream areas, including the Chesapeake Bay
- Maintaining or improving stream-bank stability on my property

Protecting Property against Damage or Loss

- Keeping vegetation from encroaching on fields or fences
- Minimizing the potential for flood damage to lands or buildings
- Discouraging pests (deer, woodchucks, snakes, insects, etc.)
- Initial costs, maintenance costs, or loss of income

Source: Adapted from tab. 3.4 in Daniel D. Dutcher, "Landowner Perceptions of Protecting and Establishing Riparian Forests in Central Pennsylvania" (Ph.D. diss., Pennsylvania State University, May 2000), 64.

of streams.[28] He asked landowners to rate the importance of twelve items thought to affect the willingness of landowners to create and maintain riparian buffers. He wanted to know whether the attitudes could be grouped into distinct dimensions that could be used as summary variables instead of using each of the eleven items separately. Using factor analysis, he found that the items factored into three dimensions. These dimensions and the items included in each dimension are listed in table 5-6. The first dimension, which he labeled "maintaining property aesthetics," included items such as maintaining a view of the stream, neatness, and maintaining open space. A second dimension contained items related to concern over water quality. The third dimension related to protecting property against damage or loss.

Factor analysis is just one of many techniques developed to explore the dimensionality of measures and to construct multi-item scales. The readings listed at the end of this chapter include some resources for students who are especially interested in this aspect of variable measurement.

Through indexes and scales, researchers attempt to enhance both the accuracy and the precision of their measures. Although these multi-item measures have received most use in attitude research, they are often useful in other endeavors as well. Both indexes and scales require researchers to make decisions regarding the selection of individual items and the way in which the scores on those items will be combined to produce more useful measures of political phenomena.

Conclusion

To a large extent, a research project is only as good as the measurements that are developed and used in it. Inaccurate measurements will interfere with the testing of scientific explanations for political phenomena and may lead to erroneous conclusions. Imprecise measurements will limit the extent of the comparisons that can be made between observations and the precision of the knowledge that results from empirical research.

Despite the importance of good measurement, political science researchers often find that their measurement schemes are of uncertain accuracy and precision. Abstract concepts are difficult to measure in a valid way, and the practical constraints of time and money often jeopardize the reliability and precision of measurements. The quality of a researcher's measurements makes an important contribution to the results of his or her empirical research and should not be lightly or routinely sacrificed.

[28] Daniel D. Dutcher, "Landowner Perceptions of Protecting and Establishing Riparian Forests in Central Pennsylvania" (PhD. diss., Pennsylvania State University, 2000).

Sometimes the accuracy of measurements may be enhanced through the use of multi-item measures. With indexes and scales, researchers select multiple indicators of a phenomenon, assign scores to each of these indicators, and combine those scores into a summary measure. Although these methods have been used most frequently in attitude research, they can also be used in other situations to improve the accuracy and precision of single-item measures.

TERMS INTRODUCED

Alternative-form method. A method of calculating reliability by repeating different but equivalent measures at two or more points in time.

Bias. A type of measurement error that results in 'systematically over- or under-measuring the value of a concept.

Construct validity. Validity demonstrated for a measure by showing that it is related to the measure of another concept.

Content validity. Validity demonstrated by ensuring that the full domain of a concept is measured.

Convergent construct validity. Validity demonstrated by showing that the measure of a concept is related to the measure of another, related concept.

Correlation matrix. A table showing the relationships among discrete measures.

Dichotomous variable. A nominal-level variable having only two categories that for certain analytical purposes can be treated as a quantitative variable.

Discriminant construct validity. Validity demonstrated by showing that the measure of a concept has a low correlation with the measure of another concept that is thought to be unrelated.

Face validity. Validity asserted by arguing that a measure corresponds closely to the concept it is designed to measure.

Factor analysis. A statistical technique useful in the construction of multi-item scales to measure abstract concepts.

Guttman scale. A multi-item measure in which respondents are presented with increasingly difficult measures of approval for an attitude.

Interitem association. A test of the extent to which the scores of several items, each thought to measure the same concept, are the same. Results are displayed in a correlation matrix.

Interval measurement. A measure for which a one-unit difference in scores is the same throughout the range of the measure.

Level of measurement. The extent or degree to which the values of variables can be compared and mathematically manipulated.

Likert scale. A multi-item measure in which the items are selected based on their ability to discriminate between those scoring high and those scoring low on the measure.

Measurement. The process by which phenomena are observed systematically and represented by scores or numerals.

Mokken scale. A type of scaling procedure that assesses the extent to which there is order in the responses of respondents to multiple items. Similar to Guttman scaling.

Nominal measurement. A measure for which different scores represent different, but not ordered, categories.

Operational definition. The rules by which a concept is measured and scores assigned.

Operationalization. The process of assigning numerals or scores to a variable to represent the values of a concept.

Ordinal measurement. A measure for which the scores represent ordered categories that are not necessarily equidistant from each other.

Random measurement error. Errors in measurement that have no systematic direction or cause.

Ratio measurement. A measure for which the scores possess the full mathematical properties of the numbers assigned.

Reliability. The extent to which a measure yields the same results on repeated trials.

Split-halves method. A method of calculating reliability by comparing the results of two equivalent measures made at the same time.

Summation index. A multi-item measure in which individual scores on a set of items are combined to form a summary measure.

Test-retest method. A method of calculating reliability by repeating the same measure at two or more points in time.

Validity. The correspondence between a measure and the concept it is supposed to measure.

SUGGESTED READINGS

DeVellis, Robert F. *Scale Development: Theory and Applications.* 3rd ed. Thousand Oaks, Calif.: Sage, 2011.

Kim, Jae-On, and Charles W. Mueller. *Introduction to Factor Analysis: What It Is and How to Do It.* A Sage University Paper: Quantitative Applications in the Social Sciences no. 07–013. Beverly Hills, Calif.: Sage, 1978.

Mertler, Craig A., and Rachel A. Vannatta. *Advanced and Multivariate Statistical Methods: Practical Application and Interpretation.* 5th ed. Glendale, Calif.: Pyrczak, 2013.

Netemeyer, Richard G., William O. Bearden, and Subhash Sharma. *Scaling Procedures: Issues and Applications.* Thousand Oaks, Calif.: Sage, 2003.

Walford, Geoffrey, Eric Tucker, and Madhu Viswanathan. *The Sage Handbook of Measurement.* Los Angeles, Calif.: Sage, 2010.

CHAPTER

6 Research Design:

Making Causal Inferences

CHAPTER OBJECTIVES

6.1 Explain the ways in which causal assertions can be verified.

6.2 Identify the types and characteristics of randomized experiments.

6.3 Discuss quasi-experimental design.

6.4 Relate how natural experiments, including intervention analysis, work.

6.5 Describe intervention analysis.

6.6 Summarize different types of observational studies, such as small-*N* designs and cross-sectional designs.

6.7 Relate the benefits of times series design.

DO NEGATIVE CAMPAIGNS ENCOURAGE OR DISCOURAGE voter interest and turnout? This is really a question about causality. A causal assertion goes beyond the claim that one thing is associated with another. It asserts, instead, that one event or entity "leads to" or "produces" or "brings about" something else. Establishing causal connections is the gold standard of modern empirical political science.[1] But this is no easy task. This chapter explains the logic behind the search for causation and how to design or plan research to make legitimate causal inferences.

A **research design** is a plan that shows how one intends to study an empirical question. It indicates what specific theory or propositions will be tested, what

[1] Just a reminder: chapter 2 notes that empiricism is a widely accepted epistemological stance in the social sciences. But many practitioners believe that the quest for causal laws of political behavior is a fool's errand and that a political scientist's proper goal is understanding or interpreting phenomena.

the appropriate "units of analysis" (e.g., people, nations, states, organizations) are for the tests, what measurements or observations (that is, data) are needed, how all this information will be collected, and which analytical and statistical procedures will be used to examine the data. All the parts of a research design should work to the same end: drawing sound conclusions supported by observable evidence.

We will discuss various types of designs along with their advantages and disadvantages. Just as important, we will show how a poor research strategy can result in uninformative or misleading results. Poor planning may produce insignificant or erroneous conclusions, no matter how original and brilliant the investigator's ideas and hypotheses happen to be.

Many factors affect the choice of a design. One is the purpose of the investigation. Whether the research is intended to be exploratory, descriptive, or explanatory will most likely influence its design. The project's feasibility or practicality is another consideration. Some designs may be unethical, while others may be impossible to implement for lack of data or insufficient time and money. Researchers frequently must balance what is possible to accomplish against what would ideally be done to investigate a particular hypothesis. Consequently, many common designs entail unfortunate but necessary compromises, and thus the conclusions that may be drawn from them are more tentative and incomplete than anyone would like.

Verifying Causal Assertions

Causal versus Spurious Relationships

Let us return to the question of the effects of campaign advertising on voting. A tentative hypothesis is that negative ads, repeated over and over, bore, frustrate, and even anger potential voters to the point that they think twice about going to the polls. Consequently, we might expect that the more citizens are subjected to commercials and advertisements that vilify candidates, the less inclined they will be to vote. Therefore, in a campaign flooded with negative ads, turnout will be lower than in one in which the candidates stick to the issues. We might even be tempted to make the stronger claim that negative political advertising *causes* a decline in participation.

How could we support such assertions? Just after an election, it might be possible to interview a sample of citizens, ask them if they had heard or been aware of attack ads, and then determine whether or not they had voted. We might even find a relationship or connection between exposure and turnout. Let's say, for instance, that all those who report viewing negative commercials tell us that they did *not* vote, whereas all those who were not aware of these ads cast ballots. We might summarize the hypothetical results in a simple table. Let *X* stand for whether or not people saw the campaign ads and *Y* for whether or not they voted. (We will see the reason for using these letters in a moment.) What this table symbolizes is a relationship or association between *X* and *Y*.

This strategy, frequently called opinion research, involves an investigator observing behavior indirectly by asking people questions about what they believe and how they act. Since we do not directly observe their actions, we can only take the respondents' word about whether or not they voted or saw attack ads.

What can we make of the findings in table 6-1? Yes, there is a relationship. It is sometimes called a correlation or perhaps, less formally, an association. Note that 100 percent of the people who said they were "exposed" also said they did not vote, and vice versa for those who did not watch any ads. But does that mean that negative advertising causes a decline in turnout? After all, it is possible that those who

HELPFUL HINTS

Causality versus Correlation

The ability to tell the difference between causation and correlation is an essential skill for political scientists and interested citizens alike. Why? Because so many arguments about policy and politics contain statements that may or may not be legitimately or reasonably interpreted as causal.

In social science research as well as common parlance, a **correlation** is simply a statement that two things are systematically related. But that's the extent of the information carried by a statement of correlation.

A *causal* declaration, by contrast, communicates much more. A change in the state of one thing *brings about* (in full or in part) a change in the state of another. This statement carries with it claims about time order and the elimination of alternative explanations for the observed relationship.

Check out more Helpful Hints at **edge.sagepub.com/johnson8e**

missed the ads differ in other ways as well from those who saw them. Perhaps they have a higher level of education and *that* accounts for their higher turnout rate. Or maybe they had a generally strong sense of civic duty and would always vote, no matter what the campaigns do or say.

TABLE 6-1 **Voting Intention by Ad Exposure**

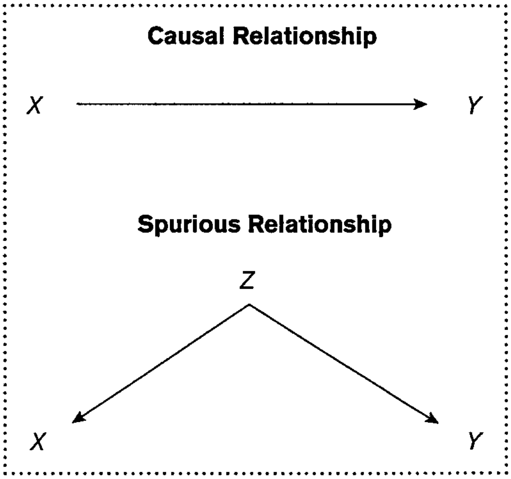
Y: Voted?	X: Yes, exposed	X: No, not exposed
Yes	0%	100%
No	100%	0%

Note: Hypothetical data.

At the same time, people with less education might watch a lot of television and *coincidentally* not bother voting in any election. If conditions of these sorts hold, we may observe a connection between advertisement exposure and turnout, but it would not be a *causal* relationship. And outlawing negative campaigning would not necessarily have any effect on turnout because the one does not cause the other. In this case, the association would be an example of what we call a *spurious,* or false, relationship.

A spurious relationship arises because two things are both affected by a third factor and thus appear to be related. Once this additional factor has been identified and controlled for, the original relationship weakens or disappears altogether. To take a trivial example, we might well find a positive relationship between the number of operations in hospitals and the number of patients who die in them. But this doesn't mean that operations cause deaths. Rather, it is probably the case that people with serious illnesses or injuries need operations *and* because of their conditions are prone to die. Figure 6-1 illustrates causal and spurious relationships.[2]

FIGURE 6-1 **Causal and Spurious Relationships**



Distinguishing real, causal relations from spurious ones is an important part of any scientific research. To explain phenomena fully, we must know how and why two things are connected, not simply that they are associated. Thus, one of the major goals in designing research is to come up with a way to make valid causal inferences. Ideally, such a design does three things:

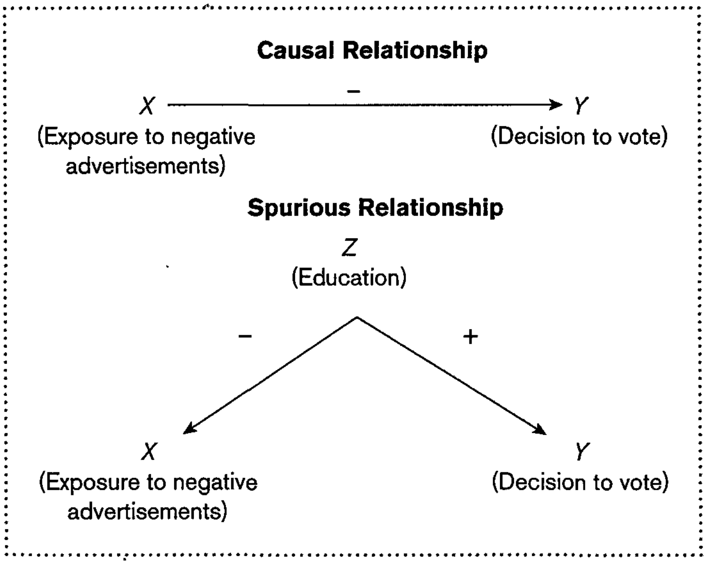
1. *Covariation:* It demonstrates that the alleged cause (call it *X*) does in fact covary with the supposed effect, *Y.* Our simple study of advertising and voting does this because, as we saw in table 6-1, viewing negative advertisements is connected to nonvoting, and not viewing the ads is associated with voting. Public opinion polls or surveys can relatively easily identify associations. To make a causal inference, however, more is needed.

[2] See chapters 13 and 14 for a more thorough discussion of spurious relationships.

2. *Time order:* The research must show that the cause preceded the effect: *X* must come before *Y* in time. After all, can an effect appear before its cause? In our survey of citizens, we might reasonably assume that the television ads preceded the decision to vote or not. But note that however reasonable this assumption may be, we have not really demonstrated it empirically. And in other observational settings, it may be difficult, if not impossible, to tell whether *X* came before or after *Y*. Still, even if we can be confident of the time order, we have to demonstrate that a third condition holds.
3. *Elimination of possible alternative causes, sometimes termed "confounding factors":* The research must be conducted in such a way that all possible joint causes of *X* and *Y* have been eliminated. To be sure that negative television advertising directly depresses turnout, we need to rule out the possibility that the two are connected by some third factor, such as education or interest in politics.

Figure 6-2 shows the necessity for the third requirement. How do you interpret these so-called causal models or arrow diagrams? The first one (Causal Relationship) shows a "true" causal connection between *X* (ad exposure) and *Y* (voting). The arrow indicates causality: *X* causes *Y*. If this is the way the world really is, then attack advertisements have a direct link to nonvoting. The arrowhead indicates the direction of causality, because in this example *X* causes *Y* and not vice versa. In the second diagram (Spurious Relationship), by contrast, the *X* and *Y* are not directly related; there is no causal arrow between them. Yet an apparent-association is produced by the action of a third factor, *Z*. Hence, the presence of the third factor, *Z* (education), creates the impression of a causal relationship between *X* and *Y*, but this impression is misleading, because once we take into account the third factor—in language we use later, "once we control for *Z*"—the original relationship weakens or disappears.

FIGURE 6-2 Models of Advertising Exposure and Voting



It might not be going too far to say that causal assertions are the life blood of political and policy discourse. Take just about any contentious subject. Its manifest argument may be about "we should . . ." statements. But underlying the argument, we guarantee, you will always find causal assertions (e.g., "We should limit greenhouse gas emissions because they cause an increase in global temperatures").

Since virtually any potential relations of interest could be spurious, how do we discern between direct and indirect linkages among variables? The answer leads to research design, because how we frame problems and plan their solutions greatly affects the confidence we can have in our results. Asking a group of people about what they have seen and heard in the media and relating their answers to their reported behavior is known in common parlance as "polling." A more formal term is *survey research*. This involves the direct or indirect collection of information from individuals by asking them questions, having them fill out forms, or other means. (We discuss survey research in chapter 10.) This approach is perhaps the most commonly used in the social sciences, and it is the one followed in the hypothetical example above. A difficulty with survey research, however, is that it is not always a reliable way to make causal inferences. For this reason, many social scientists think laboratory experiments lead to more valid conclusions.

The Classical Randomized Experiment

An **experiment** allows the researcher to control exposure to an experimental variable (often called a **test stimulus**, **test factor**, or independent variable), the assignment of subjects to different groups, and the observation or measurement of responses and behavior. Although most political scientists do conduct experiments[3]—recall Stephen Ansolabehere and his colleagues' study of the effects of campaign advertising described in chapter 1[4]— most research in the field uses nonexperimental methods. This situation results partly from the nature of the phenomena of greatest interest to political science, such as who votes in actual elections rather than in experimental settings. Nevertheless, understanding experimental design is crucial for both students of political science and citizens because it provides an especially clear way to see what must done to validate or confirm or support a causal claim.

As we noted earlier, making a valid causal claim involves showing three things: covariation, time order, and the absence of confounding factors. In theory, an experiment can unambiguously accomplish all these objectives. How? Let's

[3] For a review of experimentation in political science, see Rose McDermott, "Experimental Methods in Political Science," *Annual Review Political Science* 5, no. 1 (2002): 31–61. Available at http://www.sant.ox.ac.uk/people/knicolaidis/mcdermott.pdf

[4] Ansolabehere, Iyengar, Simon, and Valentino, "Does Attack Advertising Demobilize the Electorate?" *American Political Science Review* 88, no. 4 (1994): 829–38.

look at the following five basic characteristics of a **classical randomized experimental design**:[5]

1. The experimenter starts by establishing two groups: an **experimental group** (actually, there can be more than one), which receives or is exposed to an experimental treatment or test factor, and a **control group**, so called because its subjects do not undergo the experimental manipulation or receive the experimental treatment or test stimulus. So, for example, Ansolabehere and his colleagues had some citizens (the experimental group) watch a negative political ad and others (the control group) watch a nonpolitical commercial. The investigators determined who watched the political ad and who watched the nonpolitical commercial; they did not rely on self-reports of viewership. This control over the two groups is directly analogous to a biologist exposing some laboratory animals to a chemical and leaving others alone.
2. Equally important, the researcher *randomly* assigns individuals to the groups. The subjects do not get to decide which group they join. Random assignment to groups is called **randomization**, and it means that membership is a matter of chance, not self-selection. Moreover, if we start with a pool of subjects, random assignment ensures that at the outset, both the experimental and control groups are virtually identical in all respects. They will, in other words, contain similar proportions, or averages, of females and males; liberals, moderates, conservatives, and nonpartisans; Republicans and Democrats; political activists and nonvoters; and so on. On average, the groups will not differ in any respect.[6] Randomization, as we will see, is what makes experiments such powerful tools for making causal inferences.
3. The researcher "administers" the experimental treatment (the test factor); simply put, the experimenter determines when, where, and under what circumstances the experimental group is given the stimulus.

[5] See Donald T. Campbell and Julian C. Stanley, *Experimental and Quasi-Experimental Designs for Research* (Chicago: Rand McNally, 1966), 5–6; and Paul E. Spector, *Research Designs,* A Sage University Paper: Quantitative Applications in the Social Sciences no. 07–023 (Beverly Hills, Calif.: Sage, 1981), 24–27. Four components of an ideal experiment are identified by Kenneth D. Bailey in *Methods of Social Research* (New York: Free Press, 1978), 191.

[6] If you have trouble following this idea, imagine that you have a large can of marbles, most of which are red but a few of which are blue. Now, draw randomly from the can a single marble and put it in a box. Then draw another marble—again randomly—and put this one in a second box. Repeat this process nineteen more times. At the end, you should have two boxes of twenty marbles each. If you have selected them randomly, there should be approximately the same proportion of red and blue marbles in *each* box. If you started with a can holding 90 percent red marbles and 10 percent blue, for example, each of the two boxes should hold about eighteen red marbles and two blue ones. These may not be the exact numbers—one, say, might have three blue marbles and the other just one—but these differences will be due solely to chance and will not be statistically significant.

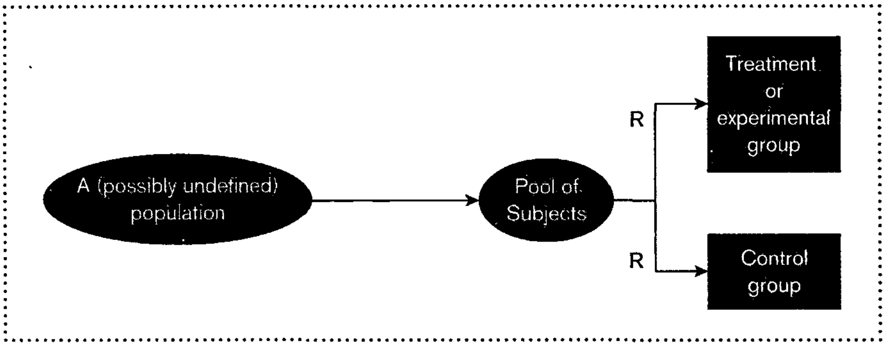
4. In an experiment, the researcher establishes a dependent variable—the response of interest—that can be measured both before and after the stimulus is given. The measurements are often called pre- and postexperimental measures, and they indicate whether or not there has been an **experimental effect**. An experimental effect, as the term suggests, reflects differences between the two groups' responses to the test factor. *This effect measures the impact of the independent variable on the dependent variable and, consequently, is a main focus of experimental research.*
5. Finally, the environment of the experiment—that is, the time, location, and other physical aspects—is under the experimenter's direction. Such control means that he or she can control or exclude extraneous factors or influences other than the independent variable that might affect the dependent variable. If, for instance, both groups are studied at the same time of day, any differences between the control and experimental subjects cannot be attributed to temporal factors.

To see how these characteristics tie in with the requirements of causal inferences, let us conduct a hypothetical randomized experiment in order to see if negative political advertising depresses the intention to vote. This case is purely hypothetical, but it roughly resembles the research conducted by Ansolabehere and his associates. More to the point, it shows the inferential power of experiments. (The example will also show some of the weaknesses of this design.)

Our hypothesis states that exposure to negative television advertising will cause people to be less inclined to vote. Stated this way, the test factor, or experimental variable, is seeing a negative ad ("yes" or "no"), and the response is the stated intention to vote ("likely" or "not likely"). We recruit a pool of subjects and randomly assign them to either an experimental (or treatment) group or a control group. It is crucial that we make the assignments randomly. We do not, for example, want to put mostly women in one group and men in the other, because if afterward we find a difference in propensity to vote, we would not be able to tell if it arose because of the advertisement or because of gender. We illustrate the procedure in figure 6-3.

Note that we draw subjects from some population, perhaps by advertising in a newspaper or giving extra credit in an American government class. This pool of subjects does not constitute a random sample of any population. After all, the subjects volunteered to participate; we did not randomly pick them. But, and here is the key, once we have a pool of individuals, we can *then* randomly assign them to the groups. Assume the first subject arrives at the test site. We could flip a coin and, depending on the result, assign him to the experimental group or to the control section. When the next person arrives, we flip the coin again and, based on just that result, send her to one or the other of the groups. If our pool consists of one hundred potential subjects, our coin tossing should result in about fifty in each group.

FIGURE 6-3 **Logic of Randomized Controlled Experiment**


A (possibly undefined) population
Pool of Subjects
R
R
Treatment or experimental group
Control group

Note: R = random assignment.

Now suppose we administer a questionnaire to the members of both groups in which we ask about demographic characteristics (e.g., age, gender, family income, years of education, place of birth) and about political beliefs and opinions (e.g., party identification, attitude toward gun control, ideology, knowledge of politics). Of course, we would also ask about the dependent variable, the intention to vote. If we compare the groups' averages on the variables, we should find that they are about the same. The experimental group may consist of 45 percent males, be on average 33.5 years old, and generally (75 percent, say) not care much for liberals. But the control group should also reflect these characteristics and tendencies. There may be only 40 percent males and the average age may be 35.0 years; but the differences reflect only chance (or, as we see in chapter 7, "sampling error"). Of greatest importance, the proportions on the response variable, the intention to vote in the next election, should be approximately the same. Thus, at the beginning of the experiment, we have two nearly identical groups.

After the initial measurement of the dependent variable (the **pretest**), we start the experiment. To disguise our purpose, we tell the informants that we are interested in television news. Those assigned to the experimental treatment go to room 101, those in the control panel to room 106. Both groups now watch an identical fifteen-minute news broadcast. So far, both groups have been treated the same. If there are any differences between them, they are the result of happenstance.

Next, the first set of subjects sees a thirty-second negative political ad, which we have constructed to be as realistic as possible, while the others see a thirty-second commercial about hair conditioners, also as true to life as we can make it. The different treatment constitutes the experimental manipulation (seeing versus not seeing a negative political advertisement). After the commercials have aired, we show both groups another fifteen-minute news clip.

TABLE 6-2 **Results of Hypothetical Media Experiment**

Group	Before measure of intention (% intending to vote)	After measure of intention (% intending to vote)
Experimental	65	45
Control	63	62

Note: Hypothetical experimental data

When the broadcast is over, we wrap up the experiment by administering parts of the first questionnaire again and measuring the likelihood of voting. This calculation gives us an indication of the size of the experimental variable's effect, if there is one. Hypothetical results from this experiment are shown in table 6-2.

Note, first, that both control and experimental subjects had about the same initial stated intention of voting (63 and 65 percent, respectively), as we would expect, because the participants had been randomized. But the posttest measurement shows quite a change for the experimental group: the percentage intending to vote has dropped from 65 to 45 percent. But the control group has hardly changed at all. The treatment effect on the experimental group is 65% – 45% = 20%, quite a reduction in civic-mindedness if it turns out to be valid in the general population (see the following discussion).

So we might conclude that the experimental factor did indeed cause a decline in intention to vote. How can we make this inference? In the first place, the experimental design satisfied all the conditions necessary for making such claims. In table 6-2, we show that the two variables, exposure to negative ads and intention to vote, covary; those who have seen a negative ad are much less likely to vote than are those who did not (45% versus 62%). We have also established the time order, since we explicitly determined the timing of the experimental treatment and the subsequent posttest measurement. Finally, and most convincing of all, we have been able to rule out any reasonable alternative explanation for the covariation, for our randomization and experimental manipulation ensured that the groups differed *only* because one received the treatment and the other did not. Since that was the only difference, the gap between the posttest percentages of the two groups could be attributed only to viewing the commercial.

The purpose of an experiment is to isolate and measure the effects of the independent variable on a response. Researchers want to be able to separate the effect of the independent variables from the effects of other factors that might also influence the dependent variable. Control over the random assignment of subjects to experimental and control groups is the key feature of experiments, because it helps them to "exclude," rule out, or control for the effects of factors that might create a spurious relationship.

The Power of Random Assignment

As we have stressed, the way researchers actually assign subjects to control and experimental groups is important. The best way is randomization, or assigning subjects to groups not according to one of their characteristics such as gender or age, under the assumption that extraneous factors will affect all groups equally and thus "cancel out." Random assignment is an especially attractive choice when it is not possible to specify possible extraneous factors in advance or when there are so many that assigning subjects to experimental and control groups in a manner that ensures the equal distribution of these factors is not possible.

Even given random assignment, extraneous factors may not be totally randomly distributed and, therefore, can affect the outcome of the experiment. This is especially likely if the number of subjects is small. Prudent researchers do not assume that all significant factors are randomly distributed just because the study design has randomized the study's participants. So, in addition to random assignment, investigators use pretests to see if the control and experimental groups are, in fact, equivalent with regard to those factors that are known to influence the outcome or suspected of doing so.

One of the biggest obstacles to experimentation in social science research is the inability of researchers to control the assignment of subjects to experimental and control groups. Even though the point of conducting an experiment is to test whether a treatment or program has a beneficial effect, it is often practically, ethically, or politically difficult to assign subjects to different treatment and control groups. Suppose, for example, that one group was to receive a generous welfare package while another one got nothing. How long could such an experiment go on? Most readers, we hope, have followed the logic of our arguments. But they must be flabbergasted at the unrealism of the hypothetical example we introduced and wonder how anyone could make a definitive statement about negative advertising based on these data, even if we had actually carried out this experiment on real people using real television commercials. Someone might exclaim, "This test is invalid!" It may be, but before jumping to that conclusion, we need to consider carefully and closely the term *validity*.

Internal Validity

If we look at causation in a particular way, statistical theory—and common sense—tell us that experiments properly conducted can lead to valid inferences about causality. In this context, however, *validity* has a particular meaning—namely, that the manipulation of the experimental or independent variable itself, and not some other variable, *did in fact* bring about the observed effect on the dependent variable. Social scientists call this kind of validity "internal validity." **Internal validity** means that the research procedure demonstrated a true cause-and-effect relationship that was not created by spurious factors. Social scientists generally believe that the type of research design we have been discussing—a randomized controlled experiment—has strong internal validity. But it is not foolproof.

Several things can affect a research study's internal validity. As we have argued, the principle strength of experimental research is that the researcher has enough control over the environment to make sure that exposure to the experimental stimulus is the only significant difference between experimental and control groups. However, sometimes *history,* or events other than the experimental stimulus that occur between the pretest and posttest measurements of the dependent variable, will affect the dependent variable. For example, suppose that after being selected and assigned to a room, the experimental subjects happen to hear a radio program that undercuts their faith in the electoral process. Such a possibility might arise if there was a long lag between the first measurement of their attitudes and the start of the experiment.

Another potential confounding influence is *maturation,* or a change in subjects over time that might produce differences between experimental and control groups. For example, subjects may become tired, confused, distracted, or bored during the course of an experiment. These changes may affect their reaction to the test stimulus and introduce an unanticipated effect on posttreatment scores.

Test-subject interaction, the process of measuring the dependent variable prior to the experimental stimulus, may itself affect the posttreatment scores of subjects. For example, simply asking individuals about politics on a pretest may alert them to the purposes of the experiment. And that, in turn, may cause them to behave in unanticipated ways. Similarly, suppose a researcher wanted to see if watching a presidential debate makes viewers better informed than nonviewers. If the researcher measures the political awareness of the experimental and control groups prior to the debate, he or she runs the risk of sensitizing the subjects to certain topics or issues and contributing to a more attentive audience than would otherwise be the case. Consequently, we would not know for sure whether any increase in awareness was due to the debate, the pretest, or a combination of both. Fortunately, some research designs have been developed to separate these various effects.

Selection bias can also lead to problems. Such bias can creep into a study if subjects are picked (intentionally or not) according to some criterion and not randomly. A common selection problem occurs when subjects volunteer to participate in a program. Volunteers may differ significantly from nonvolunteers; for example, they may be more compliant and eager to please, healthier, or more outgoing. Sometimes a person might be picked for participation in an experiment because of an extreme measurement (very high or very low) of the dependent variable. As we stressed, in assigning subjects to experimental and control groups, a researcher hopes that the two groups will be equivalent. If subjects selectively drop out of the study, experimental and control groups that were the same at the start may no longer be equivalent. Thus, **experimental mortality**, or the differential loss of participants from comparison groups, may raise doubts about whether the changes and variation in the dependent variable are due to manipulation of the independent variable.

Another possible problem comes from **demand characteristics**, or aspects of the research situation that cause participants to guess at the investigator's goals and adjust their behavior or opinions accordingly. You may remember from chapter 2 that the human ability to empathize and anticipate others' feelings and intentions (self-reflection) troubles some methodologists, who wonder if this trait doesn't make behavior inaccessible to scientific inquiry. Most political science experimenters don't think so, but they do realize that test subjects can interact with the experimental personnel and setting in subtle and occasionally unpredictable ways. It has been found that people often want to "help" or contribute to an investigator's goals by acting in ways that will support the main hypotheses.[7] Perhaps something about our experiment on political advertising tips off subjects that we, the researchers, expect to find that negative ads depress turnout, and perhaps they (even unconsciously) adjust their feelings in order to prove the proposition and hence please us. In this case, it is the desire to satisfy the researchers' objectives that affects the disposition to vote, not the commercials themselves. This is not a minor issue. You may have heard about "double-blind studies" in medical research. The goal of this kind of design is to disguise to both patients and attendants who is receiving a real experimental medicine and who is receiving a traditional medicine or placebo, thus reducing the possibility that demand characteristics affect the dependent variable.

In short, a lot of things can go wrong in even the most carefully planned experiment. Nevertheless, experimental research designs are better able to resist threats to internal validity than are other types of research designs. In fact, they provide an ideal against which other research strategies may be compared. Moreover, we discuss below some ways to mitigate these potential errors. Yet even if we devised the most rigorous laboratory experiment possible to test for media effects on political behavior, some readers still might not be convinced that we have found a cause-and-effect relationship that applies to the "real world." What they are concerned about, perhaps without being aware of the term, is externality validity.

[7] Martin T. Orne, "On the Social Psychology of the Psychological Experiment: With Particular Reference to Demand Characteristics and Their Implications," *American Psychologist* 17, no. 11 (1962): 776–83.

External Validity

External validity, the extent to which the results of a study can be generalized across populations, times, and settings, is the touchstone for natural and social scientists alike. Gerber and Green explained:

> When evaluating the external validity of political experiments, it is common to ask whether the stimulus used in the study resembles the stimuli of interest in the political world, whether the participants resemble the actors who are ordinarily confronted with these stimuli, whether the outcome measures resemble the actual political outcomes of theoretical or practical interest, and whether the context within which actors operate resembles the political context of interest.[8]

In short, the results of a study have "high" external validity if they hold for the world outside of the experimental situation; they have low validity if they only apply to the laboratory.

What sorts of things can compromise one's results? One possibility is that the effects may not be found using a different population. Refer again to figure 6-3, which showed that a pool of participants are selected from some population (of possibly unknown characteristics) and then assigned to one of two groups. But what if the population from which they have been drawn does not reflect any meaningful broader population? Suppose, for instance, we conducted an experiment on sophomores from a particular college. Results might be valid for second-year students attending that particular school but not for the public at large. Indeed, the conclusions might not apply to other classes at that or any other university. To take another example, findings from an experiment investigating the effects of live television coverage on legislators' behavior in state legislatures with fewer than one hundred members may not be generalizable to larger state legislative bodies or to Congress. In general, if a study population is not representative of a larger population, the ability to generalize about the larger population will be limited.

Other Randomized Experiments

Now that we have discussed the classical randomized experiment, let's consider some extensions of this approach. Each one represents a different attempt to retain experimental control over the experimental situation while at the same time dealing with threats to internal and external validity. Although you may not have an

[8] Alan S. Gerber and Donald P. Green, "Field Experiments and Natural Experiments," in *The Oxford Handbook of Political Methodology*, ed. Janet M. Box-Steffensmeir, Henry E. Brady, and David Collier (New York: Oxford University Press, 2008), 358.

opportunity to employ these designs, knowledge of them will help you understand published research and determine whether the research design employed supports the author's conclusions.

Posttest Design

A simple variant, the **posttest design**, involves two groups and two variables, one independent and one dependent, as before. Likewise, subjects are randomly assigned to one or the other of two groups. One group, the experimental group, is exposed to a treatment or stimulus, and the other, the control group, is not or is given a placebo. Then the dependent variable is measured for each group. The difference between this and the classical randomized experiment is that there is no pretest, so one cannot be certain that at the outset the two groups (experimental and control) have the same average levels on all relevant variables (see figure 6-4).

FIGURE 6-4 **Simple Posttest Experiment**

		Posttest
R Experimental Group	*X*	M_{exp}
R Control Group		$M_{control}$

X = Experimental manipulation

M = Measurements

R = Random assignment of subjects to groups

Nonetheless, researchers using this design can justifiably make causal inferences because they know that the treatment occurred prior to measurement of the dependent variable and that any difference between the two groups on the measure of the dependent variable is attributable to the difference in the treatment. Why? This design still requires random assignment of subjects to the experimental and control groups and, therefore, assumes that extraneous factors have been controlled for. It also assumes that, prior to the application of the experimental stimulus, both groups were equivalent with respect to the dependent variable. If the assignment to experimental or control groups is truly random, and the size of the two groups is large, these are ordinarily safe assumptions. However, if the assignment to groups is not truly random or the sample size is small, or both, then posttreatment differences between the two groups may be the result of pretreatment differences and not the result of the independent variable. Because it is impossible with this design to tell how much of the posttreatment difference is simply a reflection of pretreatment differences, a classical experimental research design is considered to be a stronger design.

Repeated-Measurement Design

Naturally, when an experiment uses both a pretest and a posttest, the pretest comes before the experiment starts and the posttest afterward. But exactly how long before and how long afterward? Researchers seldom know for sure. Therefore, an experiment, called a **repeated-measurement design**, may contain several pretreatment and posttreatment measures, especially when researchers don't know exactly how quickly the effect of the independent variable should be observed or when the most reliable pretest measurement of the dependent variable should be taken. An example of a repeated-measurement experimental design would be an attempt to test the relationship between watching a presidential debate and support for the candidates. Suppose we started out by conducting a classical experiment, randomly assigning some people to a group that watches a debate and others to a group that does not watch the debate. On the pre- and posttests, we might measure the scores shown in table 6-3.

TABLE 6-3 **Pre- and Posttest Scores in Non-Repeated-Measurement Experiment**

	Predebate Support for Candidate X	Treatment	Postdebate Support for Candidate X
Experimental group	60	Yes	50
Control group	55	No	50

Note: Hypothetical data.

These scores seem to indicate that the control group was slightly less supportive of Candidate X before the debate (that is, the random assignment did not work perfectly), and that the debate led to a decline in support for Candidate X of 5 percent: (60–50) – (55–50). Suppose, however, that we had the additional measures shown in table 6-4.

It appears now that support for Candidate X eroded throughout the period for both the experimental and control groups and that the rate of decline was consistently more rapid for the experimental group (that is, the two groups were not equivalent prior to the debate). Viewed from this perspective, it seems that the debate had no effect on the experimental group, since the rate of decline both before and after the debate was the same. Hence, the existence of multiple measures of the dependent variable, both before and after the introduction of the independent variable, would lead in this case to a more accurate conclusion regarding the effects of the independent variable.

Multiple-Group Design

To this point, we have discussed mainly research involving one experimental and one control group. In a **multiple-group design**, more than one experimental or control group are created so that different levels of the experimental variable can be compared. This is useful if the independent variable can assume several values or if the researcher wants to see the possible effects of manipulating the independent variable in several different ways. Multiple-group designs may involve a posttest only or both a pretest and a posttest. They may also include a time series component.

TABLE 6-4 Pre- and Posttest Scores in Repeated-Measurement Experiment

	Pretest				Posttest		
	First	Second	Third	Treatment	First	Second	Third
Experimental group	80	70	60	Yes	50	40	30
Control group	65	60	55	No	50	45	0

Note: Hypothetical data.

Here's an example. The proportion of respondents who return questionnaires in a mail survey is usually quite low. Consequently, investigators have attempted to increase response rates by including an incentive or token of appreciation inside the survey. Since incentives add to the cost of the survey, researchers want to know whether or not the incentives increase response rates and, if so, which incentives are most effective and cost-efficient. To test the effect of various incentives, we could use a multiple-group posttest design. If we wanted to test the effects of five treatments, we could randomly assign subjects to six groups. One group would receive no reward (the control group), whereas the other groups would each receive a different reward—for example, 25¢, 50¢, $1.00, a pen, or a key ring. Response rates (the posttreatment measure of the dependent variable) for the groups could then be compared. In table 6-5 we present a set of hypothetical results for such an experiment.

The experimental data indicate that rewards increase response rates and that monetary incentives have more effect than do token gifts. Furthermore, it seems that the dollar incentive is not cost-effective, since it did not yield a sufficiently greater response rate than the 50¢ reward to warrant the additional expense. Other experiments of this type could be conducted to compare the effects of other aspects of mail questionnaires, such as the use of prepaid versus promised monetary rewards or the inclusion or exclusion of a prestamped return envelope.

Randomized Field Experiments

As might be readily guessed, laboratory experiments, whatever their power for making causal inferences, cannot be used to study many, if not most, of the phenomena that interest political scientists. This is especially true for the study of public policies and programs. Imagine trying to discover whether or not a desegregation plan could ultimately lead to higher average test scores in school districts across the country. At a minimum you would need to randomly assign the integration schemes—the treatment—to a sample of school districts while using others as controls. That would be possible in theory but impossible in practice. How would you force districts to integrate? If you accept voluntary participation in place of random allocation, the voluntary districts might very well differ from those that won't accept the plan in unknown ways.

TABLE 6-5 Mail Survey Incentive Experiment

(Random Assignment)	Treatment	Response Rate (%)
Experimental Group 1	25¢	45.0
Experimental Group 2	50¢	51.0
Experimental Group 3	$1.00	52.0
Experimental Group 4	pen	38.0
Experimental Group 5	key ring	37.0
Control Group	no reward	30.2

Note: Hypothetical data.

Nonetheless, the basic principles of experimental design can be taken into the field. A **field experiment** adopts the logic of randomization and variable manipulation by applying these techniques to naturally occurring situations and units.[9] Samples of individuals or aggregates of people (e.g., students in a city's school districts) are randomly chosen to receive a treatment (e.g., a new mathematics curriculum), while others receive another or are used as controls. Once the experiment has been concluded, the investigator can take posttest measurements to determine if the treatment had an effect. Let's look at an example.

[9] For an overview, see Thomas D. Cook and William R. Shadish, "Social Experiments: Some Developments over the Past Fifteen Years," *Annual Review of Psychology* 45, no. 1 (1994): 545–80.

Like others mentioned in this chapter, David Niven speculated about the effects of campaigning on electoral participation. He first noted that

> Various . . . studies inquire about intentions to vote, or candidate preferences, but none is equipped to measure actual resulting behavior. . . . Regardless of the rigor of the researchers or the ingenious nature of their design, the laboratory remains a difficult setting in which to demonstrate the effect of negative advertising on the real world behavior of turning out to vote.[10]

As a way around this inferential obstacle, he conducted an experiment on citizens of West Palm Beach, Florida:

> Voters in the sample were randomly assigned to either the control group (700 voters who would not receive any mailings) or to one of seven experimental groups (which varied in the number of negative mailings each would receive). . . . Subjects receiving the treatment were randomly assigned to one of seven groups which received either one, two, or three negative ads. . . . After the ads were distributed and the election had occurred, official voting records were consulted to determine who cast a ballot in the election.[11]

Niven found a positive effect: turnout among the residents receiving the negative mailings was a bit higher (32.4%) than among those in the treatment condition (26.6%).[12] Hence, Niven comes down on the side of those who feel negative advertising might have a beneficial impact on voters. (Incidentally, we see again the necessity of replication and verification in empirical political science.[13])

Another, perhaps more common application of randomized field experiments is found in policy evaluation studies. **Policy evaluation** (sometimes called "policy analysis") simply means objectively analyzing the economic, political, cultural, or social impacts of public policies.[14] Targets of these sorts of research projects span policy domains from housing to health care, transportation to education, crime prevention to recycling. Occasionally, governments mandate these efficacy studies to see if the taxpayers' dollars are having a genuine effect.

[10] David Niven, "A Field Experiment on the Effects of Negative Campaign Mail on Voter Turnout in a Municipal Election," *Political Research Quarterly* 59, no. 2 (2006): 204.

[11] Ibid., 206.

[12] Ibid., 207.

[13] For similar studies, see Ted Brader, "Striking a Responsive Chord: How Political Ads Motivate and Persuade Voters by Appealing to Emotions," *American Journal of Political Science* 49, no. 2 (2005): 388–405, available at http://www.uvm.edu/~dguber/POLS234/articles/brader.pdf; David Dreyer Lassen, "The Effect of Information on Voter Turnout: Evidence from a Natural Experiment," *American Journal of Political Science* 49, no. 1 (2005): 103–18; and David W. Nickerson, Ryan D. Friedrichs, and David C. King, "Partisan Mobilization Campaigns in the Field: Results from a Statewide Turnout Experiment in Michigan," *Political Research Quarterly* 59, no. 1 (2006): 85–97, available at http://www.nd.edu/~dnickers/papers/PartyMobilization.pdf

[14] Note that policy evaluation involves much more than field experiments.

A famous field experiment was the New Jersey Income Maintenance study, funded by the Office of Economic Opportunity, which was conducted from 1967 to 1971.[15] This effort was the forerunner of other large-scale social experiments designed to test the effects of new social programs. The experiment also provides insights into the difficulty of testing the effects of public policies on a large scale in a natural setting.

Nonrandomized Designs: Quasi-Experiments

Suppose we set up an experiment like the one exploring the effects of negative campaign ads on intention to vote, but we do *not* randomly assign the students to the experimental and control groups. Instead, we use our judgment or, more likely, preexisting groups—perhaps two sections of Political Science 101. Suppose, for instance, both classes are taught in similar case-study rooms, but in one there is a monitor in front of every student, whereas the second has a single screen at the head of the room. We might decide that the treatment should be applied to the former room because students have monitors right in front of them and we can be sure that each person has a clear field of view. Those in the smaller, less-equipped room will be seeing bland commercials and can make do with a single screen. Since we are going to measure intention to vote both before and after the experiment, we reason that under the circumstances, this plan provides a reasonable approximation of a classical experiment.

More realistically, perhaps, units get picked for study because they have or are going to undergo some treatment. (Suppose we discovered that the two sections of Political Science 101 were to use different texts and Web materials in such a way that they could serve as rough approximations of our desired experimental and control groups.) Since their selection is totally independent of the investigator, he or she is merely an observer, albeit one who puts the data in a logically coherent form so they resemble experimental results. Consequently, although a quasi-experiment "looks" like a classical experiment, there is one key difference: no randomization. A **quasi-experimental design** contains treatment and control groups, but the experimenter does randomly assign individual units to these groups. The effects, if any, of putative treatments have to be inferred without the help of strong internal validity. To compensate for the lack of randomization, experimenters turn to judgment, theory, common sense, and statistical and mathematical tools to rule out spurious or confounding causes. Any scientific activity requires some inference, but as one moves from randomized designs to quasi-experiments, inferences about

[15] David Kershaw and Jerilyn Fair, *The New Jersey Income-Maintenance Experiment: Vol. 1, Operations, Surveys, and Administration* (New York: Academic Press, 1976). Also see Joseph A. Pechman and P. Michael Timpane, eds., *Work Incentives and Income Guarantees: The New Jersey Negative Income Tax Experiment* (Washington, D.C.: Brookings Institution, 1975), esp. chaps. 2 and 3.

causal effects demand more and more of the researcher's substantive knowledge and analytic skills.

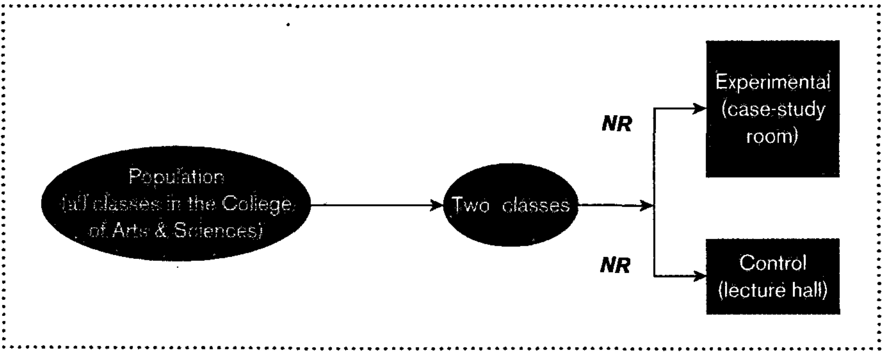
Figure 6-5 applies a quasi-experimental design to our hypothetical experiment on the effects of campaign advertising. (For simplicity, assume all the students are eligible to vote.)

In the current example, we are not adhering to the standard protocol—neither the students nor the treatment have been randomized—so our inferences could well be misleading or wrong. Why? Because we cannot be assured that at the outset the two groups are homogenous (have the same average values on all background variables). Put aside the fact that the experimental settings differ: here, the study would be conducted in dissimilar classrooms taught by possibly different instructors, a setup that violates the assumption that the groups only differ with respect to the stimulus and not with regard to the experimental conditions. Consider instead that the members of the seminar section might on average be older (or younger) or have higher (or lower) GPAs or are more (or less) interested and knowledgeable about campaigns and elections. If we find an ostensible or apparent effect—ads reduce motivation to vote—it might be because the ads really do have an effect or because of the operation of some unmeasured or unobserved factor, or both. The problem is that we just don't know. The internal validity of an experiment is suspect.

What is to be done?

Because physically assigning groups or subjects to different treatments by randomization may be difficult or impossible, social scientists turn to observational and statistical methods. These, too, follow a pattern, but their missing ingredient is randomization. This fact makes causal inferences much more tenuous. Although it

FIGURE 6-5 **Example Quasi-Experiment**


Population
(all classes in the College of Arts & Sciences)
Two classes
NR
NR
Experimental (case-study room)
Control (lecture hall)

Note: *NR* = random assignment to group.

may be stretching things a bit, we might term any such study a quasi-experiment because the goal is the same—that is, to identify causal relationships—but it does not fulfill a key requirement of experimentation: random assignment of units or treatments.

The logic can be outlined this way:

- Let Y, the dependent "variable," stand for the phenomenon of interest, the *explanandum*, or "that which must be explained."
- Identify X—a "treatment" or independent variable with at least two values or states and possibly more—that might causally affect Y.
- Look for covariation: Do values of X vary with different outcomes, or different values of Y?
- Observe the dynamics of the interaction: Did changes in X seem to precede changes in Y?

Look for and ensure that all other possible effects on Y have been taken into account. In terms of the trade, this procedure is called "controlling" or "holding constant" variables.

All methods considered in this section fit under this scheme. They look for X-Y relationships. Note that X and Y do not necessarily refer to quantitative or numerical variables. Y, for example, might consist of two categories ("war occurred" and "war did not occur") or possibly three states (peace, tension but no armed conflict, armed conflict). But beyond just looking for a relationship, the analyst factors in variables or conditions that might also cause Y. How? By judgment, careful observation, application of previous research, common sense and logic, and—in some cases, where appropriate—statistical adjustment. Throughout *Political Science Research Methods* we show how some of these techniques or methods actually work.

In our contrived example, we might have access to course rosters and the instructors' records and be able to measure the average class level, gender, major, and so forth. We hope that there would be no appreciable variation between groups on these indicators, thereby increasing our confidence that treatment did indeed have the hypothesized effect (see table 6-6).

We see that at the start, both classes had roughly the same percentages on all the variables we were able to measure explicitly, including "intention to vote" in the next election. After the quasi-experiment, the dependent variable (Y) has decreased 10 points from 60 to 50 percent in the experimental group, while it had changed hardly at all among the control subjects (55 to 52 percent). As expected, the other variable averages have stayed the same. Consequently, we make two tentative conclusions: (1) the treatment (exposure to negative ads) is *associated with* a drop in voting, and (2) this decrease is *not* explained by changes in any other measured

TABLE 6-6 Results from Hypothetical Quasi-Experiment

	Before Measurements	Postmeasurements
Experimental (negative ads in case-study room)		
Percentage intending to vote	60	50
Percentage male	45	45
Percentage liberal arts majors	70	70
Percentage undeclared	40	40
Control (bland ads in lecture hall)		
Percentage intending to vote	55	52
Percentage male	50	50
Percentage liberal arts majors	65	65
Percentage undeclared	35	35

variables, which have remained at their same levels. So, for example, the drop in expected turnout in the experimental group is presumably not due to changes in major or gender, which are constant. That leaves among the measured factors only the exposure-nonexposure difference.

Naturally, this is not a very compelling example. But it reveals the logic behind most empirical studies in political science, whether or not they are quantitative: they use judgment and explicit measurement and controls to rule out the possibility that the treatment-effect relationship is not spurious. Even if quasi-designs do not meet the standards of randomized experiments, they constantly lead to new knowledge.

Natural Experiments

If you think about it, you can take the idea of observing the effects of treatment and variables into the world. This leads to the idea of a **natural experiment**. In such studies "nature"—forces not under the investigator's control—assign individuals or units to "treatment" and "control" groups. That is, researchers observe but do not themselves manipulate the operation of the "experimental" factor. Actually, this process simply amounts to comparing groups that have been created independently.

Here's a simple example. Professor of geography Jared Diamond employed a "natural experiment of borders" to explain why two countries sharing the same island

and thus having roughly similar physical environments have drastically different standards of living. Diamond compared Haiti and the Dominican Republic, which occupy the island of Hispaniola (see table 6-7). In essence, he treated the more or less artificial and arbitrary border between the countries as a kind of ongoing experimental treatment.[16] As Diamond explained, this kind of analysis

> examines the effects of drawing a border where previously there was none . . . or the effect of removing a border where previously there was one. . . . These comparisons can shed light on the effects of differing institutions and histories. They reduce the effects of other variables, either by comparison of the same geographic area before and after the creation or removal of the border, or by simultaneous comparison of two neighboring and geographically similar areas.[17]

Until the early twentieth century, Haiti was far more populous and richer than its neighbor. Then their fortunes reversed: Haiti remains one of the world's poorest countries, while its neighbor to the east has experienced a growing economy. The data in table 6-7 point to a few of the differences. Why the turnabout in status? Diamond cited several possible causative factors:

- Precipitation and agriculture: The Dominican Republic receives most of the seasonal rainfall and hence has an agricultural advantage. Also, decades of deforestation in Haiti with attendant soil erosion eventually gave this country a disadvantage.[18]
- History: The colonial histories of the two countries differed in ways that helped produce the situation we see today. The French colonized the western (Haitian) part of the island, where they established a thriving timber and sugar export economy based on slave labor. The slaves won their freedom in the world's first and only successful slave revolt. But this success had unintended consequences. After violently driving the French out and killing most of the remaining white settlers in 1803, Haitians were suspicious of Europeans and became increasingly isolated. The Dominican Republic, by contrast, followed a more or less peaceful path to independence and early on established commercial relations with other countries. This experience led to a more open society that was not as adverse to innovation and outside ideas as its neighbor.

[16] Jared Diamond, "Intra-Island and Inter-Island Comparisons," in *Natural Experiments of History,* ed. Jared Diamond and James A. Robinson (Cambridge, Mass.: Belknap Press of Harvard University Press, 2010), 120–41.

[17] Ibid., 120–21.

[18] As Diamond explained, "even if the human societies of Haiti and the Dominican Republic had been culturally, economically, and politically identical (which they have not been), the Haitian part of Hispaniola would still have faced serious environmental problems" (ibid.).

TABLE 6-7 **Haiti and Dominican Republic Compared**

Indicators of Well-Being	Dominican Republic	Haiti
Life expectancy at birth (years)	72.4	61.0
Adult literacy rate (% ages 15 and above)	89.10	62.10
GDP per capita (USD)	6,706	1,155
Probability of not surviving to age 40 (%)	9.40	18.50

Source: Jared Diamond, "Intra-Island and Inter-Island Comparisons," in *Natural Experiments of History,* ed. Jared Diamond and James A. Robinson (Cambridge, Mass.: Belknap Press of Harvard University Press, 2010): 120–41.

- Language: Dominicans adopted the language of the colonial power, Spanish. But as Diamond noted, "Haitian slaves, who came from many different African-language groups, developed for communication a Creole language of their own. . . . Today, about 90% of Haiti's population still speaks only Haitian Creole (a language spoken by virtually no one else in the world except emigrant Haitians). . . . [Consequently,] Haitians are linguistically isolated from the rest of the world."[19]
- Political leadership: The two nations entered the modern age under two vastly different political leadership styles. By the 1930s, both were governed by tyrannical dictatorships. The Dominican Republic's General Rafael Trujillo, who ruled from 1930 to 1961, governed with an iron fist but (mostly for personal aggrandizement) developed export industries, attracted foreign investment, preserved forests, and encouraged immigration. Consequently, the economy grew under the "evil" Trujillo and his successors to the point where it could sustain a middle class and nascent democratic institutions. The story in Haiti, though, turned out much differently. The secretive François ("Papa Doc") Duvalier, Haiti's ruler from 1957 to 1971, "had little interest in economic development, export industries, or logging . . . did not bring in foreign consultants, and allowed deforestation to continue."[20] Under his and his son's leadership, the country began to lag further and further behind the Dominican Republic.

In the language of experimentation, Haiti was "exposed" to one set of levels of the "treatments" (the *X*s or explanatory factors); the Dominican Republic to another set.

[19] Ibid., 125.

[20] Ibid., 128.

TABLE 6-8 **Natural Experiment "Results"**

	Environmental Variables (X)		Cultural and Political Variables			Outcome (Y)
	Favorable climate (e.g., rainfall)	Fertile soil	History of slavery	"Universal" language	Leadership	"Successful" development
Dominican Republic	Yes	Yes	No	Yes	Yes	Yes
Haiti	No	No	Yes	No	No	No

Source: Based on Jared Diamond, "Intra-Island and Inter-Island Comparisons," in *Natural Experiments of History*, ed. Jared Diamond and James A. Robinson (Cambridge, Mass.: Belknap Press of Harvard University Press, 2010): 120–41. Available at http://www.vermontriverconservancy.org/about-vermont-river-conservancy/vrc-staff-and-board/dan-dutcher

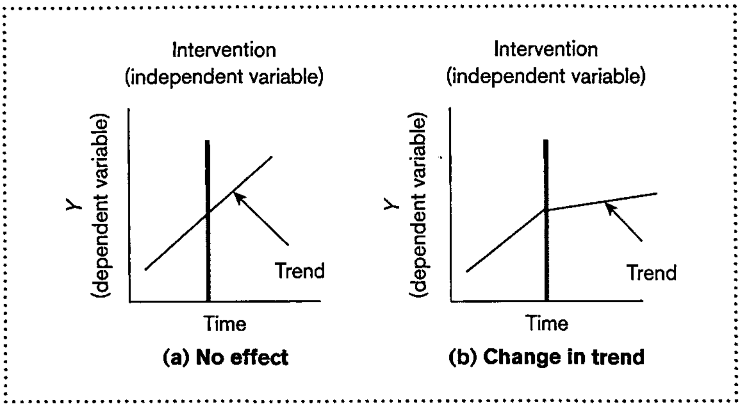
The "response" variable is, loosely, economic and political development (Y). *Successful* in this case means being higher on just about every imaginable indicator of well-being from freedom to food to health to democracy to life expectancy to political stability. Diamond hypothesized that the *X*s caused *Y*. Although the study did not explicitly present a formal analysis, the results can be represented schematically in a table (see table 6-8). We have described the research as being a quasi- or natural experiment; the logic is actually the same as that of many comparative studies, a point we demonstrate shortly.[21]

Intervention Analysis

In one version of a nonexperimental time series design, called **intervention analysis** or "interrupted time series analysis," measurements of a dependent variable are taken both before and after the "introduction" of an independent variable. Here we speak figuratively: as with the other nonrandomized designs, the occurrence of the independent variable is *observed,* not literally introduced or administered. We could observe, for instance, the annual poverty rate both before and after the ascension of a leftist party to see if regime change makes any difference on living standards. The premeasurements allow a researcher to establish trends in the dependent variable that are presumably unaffected by the independent variable so that appropriate conclusions can be drawn about posttreatment measures. Refer to figure 6-6. Panel (a) shows an increase in a dependent variable over time. (Suppose it is the poverty rate in metropolitan areas.) At a specific

[21] Incidentally, Diamond's approach embodies the principles of John Stuart Mill's (1850) "method of comparison," which we explain later.

FIGURE 6-6 **Some Possible Effects of an Intervention**



moment or period, an intervention takes place (perhaps the enactment of a jobs-training program). But the trend line remains undisturbed: *Y* grows at the same rate before and after the "appearance" of the independent variable. In this case, the intervention did not interrupt or alter the trend. (We would conclude, for example, that the program did not affect the increase in poverty.) Now consider the second figure (b). It shows an increase in *Y* until the intervention occurs, at which point the growth in the trend begins to abate. In this instance, the introduction of the factor appears to have caused the trend to flatten (e.g., the advent of job training slowed the growth in poverty).

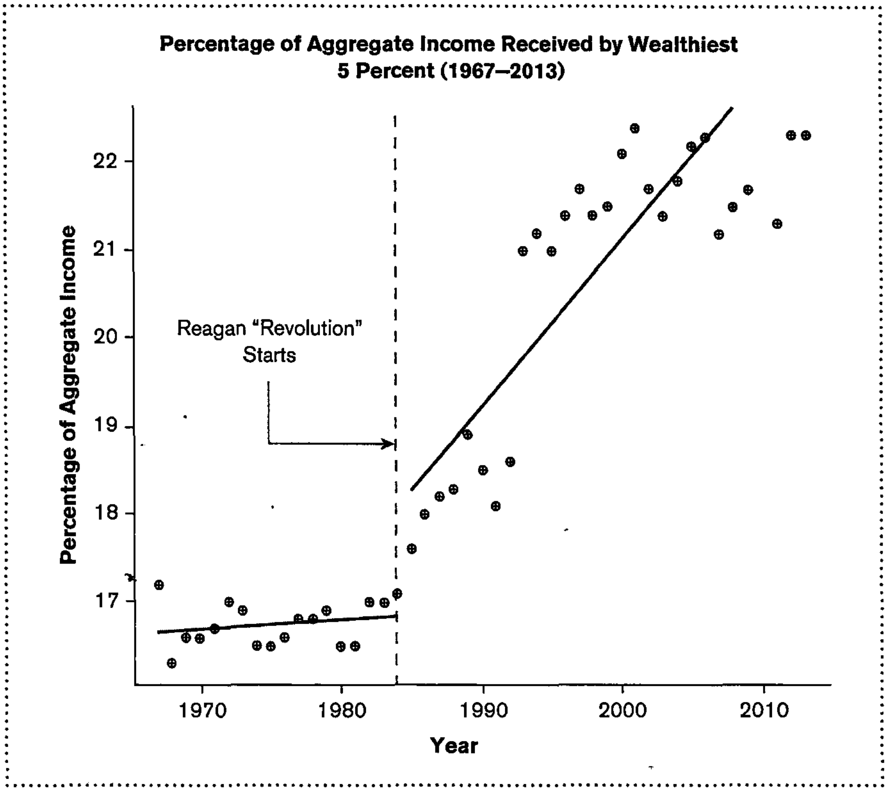
For a perhaps more realistic example, let us return briefly to the case Hacker and Pierson made about the growth of business power and income inequality in the United States (see chapter 1).[22] Recall that the authors first documented an increase in the share of America's wealth going to the wealthiest individuals. Figure 6-7 shows, for instance, that the richest 5 percent of citizens received about 16.5 percent of aggregate income in 1967; that portion had grown to nearly 23 percent by 2009. Hacker and Pierson claimed that this phenomenon—the rich getting richer, leaving less for the middle and lower classes—does not result from mere economic change or happenstance but follows as a direct result of policy changes (e.g., tax rates and financial deregulation). They explained that

> Policy—both what government has done and what, as a result of drift, it has failed to do—has played an absolutely central role in the rise of winner-take-all economic outcomes. . . . Moreover, in the main areas where the role of government appears most significant, we see a consistent pattern: active, persistent, and consequential action on the part of organized interests that stood to gain from a transformation of government's role in the American economy. A winner-take-all politics accompanied, and helped produce a winner-take-all economy.[23]

[22] Jacob S. Hacker and Paul Pierson, "Winner-Take-All Politics: Public Policy, Political Organization, and the Precipitous Rise of Top Incomes in the United States," *Politics & Society* 38, no. 2 (2010): 152–204. Available at http://pas.sagepub.com/content/38/2/152.full.pdf

[23] Ibid., 196. See also Joseph E. Stiglitz, *The Price of Inequality* (New York: Norton, 2012), 82.

FIGURE 6-7 Example Intervention Analysis



Source: US Census Bureau, *Current Population Survey: Annual Social and Economic Supplements; Historical Income Tables—Income Inequality,* table H-2: Share of Aggregate Income Received by Each Fifth and Top 5 Percent of Households (All Races) 1967 to 2013. Available at http://www.census.gov/hhes/www/income/data/historical/inequality/

Hacker and Pierson dated the transformation from the mid-1980s, when the so-called conservative or Reagan "revolution" began. Policy shifts advantageous to the wealthy, however, have been sustained with the help of Democrats in the White House and Congress.[24] It is a stretch, but we could analyze the argument with an intervention analysis. Figure 6-7 shows the share of aggregate income going to the richest 5 percent of Americans each year from 1967 to 2009. (These data are called a "time series.") We see that this share remained at about 16 to 20 percent through

[24] For example, "The shift toward a much more favorable tax regime for the wealthy has occurred largely through policy enactments. The bulk of these have occurred under Republican congressional majorities and Republican presidents (although often with significant Democratic support)" (ibid., 186).

the early part of the 1970s. Beginning in the late 1970s and accelerating when Ronald Reagan took office in 1981, the federal government began to cut taxes and roll back regulations of the finance industry. We might conceptualize this change as a policy "intervention." We see its "effects" in the jump in income going to the top group after about 1984, when Reagan was reelected to a second term; after climbing to above 20 percent, it has since stabilized. The straight lines—technically called "regression" lines—show that changes were relatively modest from 1967 to 1984 and then soared. There is, in statistical language, a shift in both the level (average) and slope (trend) in the series of income shares.

This analysis is, of course, vastly oversimplified. But it illustrates the quasi-experimental design and its pitfalls. We can imagine an omnipotent experimenter manipulating the policy regime to see what impact the change would have on incomes. The data from Hacker and Pierson's study suggest that increasing concentration of wealth had a cause—namely, the tax and other conservative policies pushed by the corporate and financial sectors. But notice our use of the weasel word *suggest*. The conclusion rests on an inference that no other factors were at work to produce the observed changes. Since no one randomized the years to control and experimental groups, we have no assurance that other unmeasured variables were at work. And Hacker and Pierson's critics argue that many other factors were surely in play. Yet, as we stress throughout the book, it is incumbent on the skeptics to identify those alternative causes and show how they, not growing business power, account for the results.

In closing this example, we point out that a "real" intervention analysis involves the application of statistical techniques to, among other things, ensure that the observed shifts in level and trend are not merely the result of chance fluctuations in the time series. A more realistic study would require measuring other variables and adjusting the data to remove the effects of random error.[25]

Before showing further designs, let us review what we have discussed regarding experimental designs. Table 6-9 compares the essential features of randomized and quasi-experiments. Randomized laboratory and field experiments enjoy a modest (and growing) place in political science's tool kit; most research relies on extensions of the quasi-experimental design. Besides logistical considerations, these approaches can lead to somewhat higher levels of external validity (realism). Still, as we have seen, *all* social and political research depends heavily on untested (sometimes unrecognized) assumptions and inferences.

[25] The canonical source is G. E. P. Box and G. C. Tiao, "Intervention Analysis with Applications to Economic and Environmental Problems," *Journal of the American Statistical Association* 70, no. 349 (1975): 70–79. Available at ftp://ftp.uic.edu/pub/depts/econ/hhstokes/e537/Box_Tiao_March_75.pdf

TABLE 6-9 Randomized and Quasi-Experiments Compared

	Design Type	
	Classical randomized experiments	Quasi-experiments
Assignment of subjects to treatment and control groups	Random	Nonrandom (judgment, self-selection, natural processes, history)
Treatment	Experimentally manipulated directly or by power of random assignment	Observation of occurrence, distribution, duration
Time order	Controlled by investigator	Inferred by investigator
Effects	Can potentially be classified as causal	Difficult to classify as "caused by" without extra-experimental data
Internal validity*	High	Medium to low
External validity*	Low	Medium
Example designs	Randomized before-and-after design Multiple-group design Field experiment Policy evaluation study	Natural experiment Comparison Intervention analysis Observational study (see table 6-10)

*For most designs.

Observational Studies

One could reasonably apply the term **observational study** to describe quasi-experimental designs in which the researcher does not manipulate experimental variables or randomly assign subjects to treatments but instead merely *observes* causal sequences and covariations. A simple comparison, for example, could be recast as a thought experiment in which the effect of a supposed treatment is examined for two or more groups. Think of "political party system" (one party, two parties, and three parties) as an experimental factor whose effects on voting turnout are of interest. (The basic hypothesis might be that turnout as a percentage of eligible voters increases as a nation moves from a one-party structure to one having two or more competitive parties.) We cannot assign a country to a party-system type, but we can compare (observe) turnout in different party systems that already exist to determine if there is at least an association between treatment and effect. Using historical and socioeconomic data, we might conclude that aside from the treatment, the countries in our sample have approximately similar values on all our measured variables. Hence, if there are differences in turnout (in the predicted direction), we might infer that they are caused by type of party system. Critics, in turn, might object that we have not included all relevant historical and political factors, that the countries differ in fundamental ways other than the type of electoral system, and that it is these unobserved, unmeasured variables that create the differences in turnout. (Of course, it is incumbent on the critic to specify some of these missing variables.) This is the sort of comparative analysis that students of political science are used to. Note, however, the underlying logic and how it differs from that of randomized studies.

We provide an overview of some the possibilities in table 6-10. When reading the table, it is important to note that many of the entries are only indicative of what can be done. Look, for example, at the row labeled "Surveys." A survey or poll usually includes anywhere from 100 to 5,000 (or more!) individuals, but polling fewer than 100 people is possible and not necessarily unsound. Furthermore, many research projects combine elements of different designs, as in a panel study with an intervention interpretation (see the following discussion). In the sections that follow, we discuss these designs in more detail.

Small-*N* Designs

CASE STUDIES AND COMPARATIVE ANALYSIS. In a **small-*N* design**, the researcher examines one or a few cases of a phenomenon in considerable detail, typically using several data collection methods, such as personal interviews, document analysis, and observation. When just one thing is under investigation, the design is often called a **case study design**; when two or more are involved, the term *comparative* or *comparative case study or analysis* is frequently used. The units of analysis or the subjects of the study can be people (e.g., prime ministers), events (e.g., outbreak of the Korean War), institutions (e.g., the US Senate), nations or alliances (e.g., NATO), decisions (e.g., passage of the Affordable Care Act), or policies (e.g., gun control legislation in Canada). The point is that one or a few cases or instances are studied in depth. As sociologist Theda Skocpol explained, these types of designs involve "too many variables and not enough cases,"[26] meaning that the investigator collects lots of data on one or a few units.

A small-*N* study may be used for exploratory, descriptive, or explanatory purposes. Exploratory case studies are sometimes conducted when little is known about a phenomenon. Researchers initially may observe only one or a few cases of that phenomenon, and careful observation of this small set of cases may suggest possible general explanations for the behavior or attributes that are observed. These explanations—in the form of hypotheses—can then be tested more systematically by observing more cases (see figure 6-8). Carefully scrutinizing the origins

[26] Theda Skocpol, *States and Social Revolutions* (New York: Cambridge University Press, 1979), 36.

TABLE 6-10 Some Observational and Statistical Designs

Design	Typical number of units or cases (*N*)[*]	Examples of units of analysis	Purpose	Examples
Small-*N* Designs				
Single case (case study)	*N* = 1	Event, nation, group, county, individual . . .[†]	Provide a detailed description and explanation.	Study of the 2014–2015 Louisiana senate election; study of the passage of the ACA (Obamacare)
Comparative[▾]	2–20	Events, nations, groups, counties, individuals . . .	Compare two or several units in relative detail.	Comparison of French and Russian Revolutions; study of ISIS, Boko Haram, and al-Qaeda
Focus group	10–20	Individuals	Often used in market research to probe reactions to stimuli such as commercials.	Test of campaign ad's effectiveness
Cross-Sectional Designs				
Surveys (polls)[▾]	100–5,000	Individuals	A large number of people are measured on several variables to search for (possibly causal) relationships.	Study of voting and public opinion
Aggregate data analysis[▾]	20–500	Aggregates:[◉] states, counties, cities, countries . . .	Variables are often averages or percentages of geographical areas, but the goal is to search for (possibly causal) relationships.	Study of the death penalty and crime rates in different states; study of the relationship between union strength and welfare spending in developed countries
Longitudinal (Time Series) Designs				
Trend analysis[▾]	20–300	Aggregates, individuals, cohorts . . .[+]	Measurements on same variables at different time periods to examine changes in levels.	Study of changing levels of trust in government; study of level of unemployment; study of occurrence of civil strife in Europe, 1900–2000
Panel study[▾]	200–5,000	Individuals, households, cohorts	The same units are measured at different times to investigate relationships, changes in strength of relationships, and causality.	Study of changes in opinions of President Barack Obama

Notes:

[▾] May rely heavily on statistical analysis.

[*] These numbers are merely suggestive; some designs involve fewer or more cases.

[◉] Data are usually summations or averages of aggregations of individuals (often in geographical areas), such as by median income in cities or counties.

[+] Individuals who experience the same event or experience or characteristics, such as a "birth cohort" (those people born in a specific year or period) or "event cohort" (e.g., those who first voted in 1972).

of political unrest within a single country may suggest general explanations for dissent, or following a handful of incumbent representatives when they return to their districts may suggest hypotheses relating to incumbent attributes, district settings, and incumbent-constituency relations.[27]

The purpose of a descriptive study may be to discover and describe what happened in a single or select few situations, thereby finding avenues for further research. Here, the emphasis is not on developing general explanations for what happened. Alternatively, in some situations a single case may provide a critical test of a theory.[28] Recall from chapter 2 that verification and falsification are crucial activities in science, so finding a single exception may cast doubt on a previously accepted proposition. Therefore, if you can find a well-documented instance in which a widely accepted or important proposition does not hold, you may make a significant contribution.

For years some scholars considered this approach a suspect or even inferior research strategy, partly because of its limited "sample sizes." Moreover, it might be thought that small-*N* designs are useless in causal analysis. But social scientists now recognize this type of design as a "distinctive form of empirical inquiry" and an important design for the development and evaluation of public policies as well as for developing explanations and testing theories of political phenomena.[29]

Proponents argue that a small-*N* design has some distinct advantages over experimental and cross-sectional designs for testing hypotheses under certain conditions. For example, a case study may be useful in assessing whether a statistical correlation between independent and dependent variables, discovered using a cross-sectional design with survey data (see the following discussion), is really causal.[30] By choosing a case in which the appropriate values of the independent and dependent variables are present, researchers can try to determine the timing of the introduction of the independent variable and how the independent variable actually caused the dependent variable. That is, they can learn whether there is an actual link between the variables and, therefore, can more likely offer an explanation for the statistical association. Benjamin Page and Robert Shapiro concluded their study of the statistical relationship between public opinion and public policy with numerous case studies.[31]

[27] See Richard F. Fenno Jr., *Home Style: House Members in Their Districts* (Boston: Little, Brown, 1978).

[28] Robert K. Yin, *Case Study Research: Design and Methods*, rev. ed., Applied Social Research Methods Series, vol. 5 (Newbury Park, Calif.: Sage, 1989), 47.

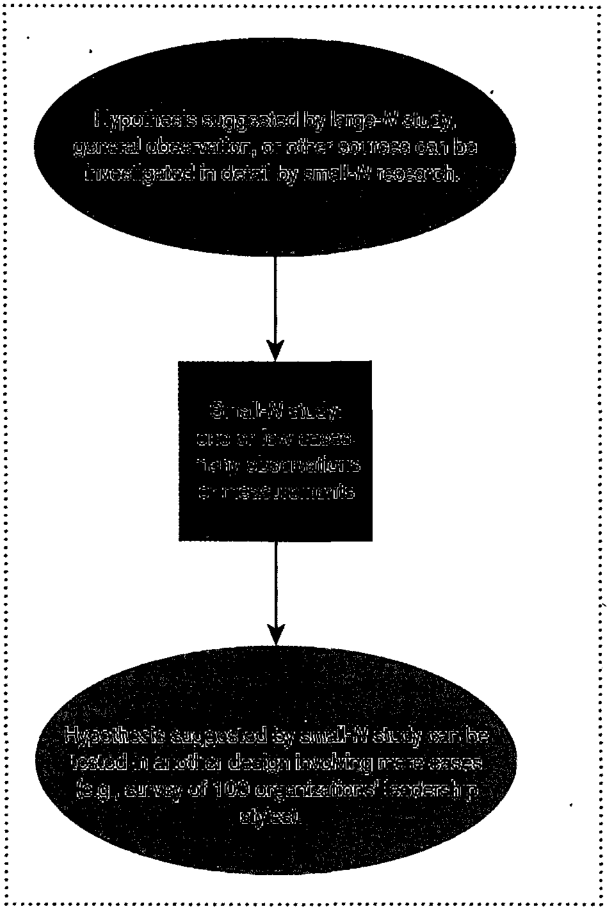
[29] Ibid., 21.

[30] Alexander L. George, "Case Studies and Theory Development: The Method of Structured, Focused Comparison," in *Diplomacy: New Approaches in History, Theory, and Policy*, ed. Paul Gordon Lauren (New York: Free Press, 1979), 46; and Skocpol, *States and Social Revolutions*, chap. 1.

[31] Benjamin I. Page and Robert Y. Shapiro, "Effects of Public Opinion on Policy," *American Political Science Review* 77, no. 1 (1983): 186.

These studies differ from experimental designs in that the researcher is able neither to assign subjects or cases to experimental and control groups nor to manipulate the independent variable. Hence the term *observational.* Yet the careful selection of a case or cases can lead to the approximation of a quasi-experimental situation. For example, a historian or political scientist may choose cases with different values of an independent variable but with the same values for important control variables. Cases with similar environments can be chosen. Furthermore, lack of complete control over the environment or context of a phenomenon can be seen as useful. If it can be shown that a theory actually works and is applicable in a real situation, then the theory may more readily be accepted. This may be especially important, for example, in testing theories underlying public policies and public programs.

FIGURE 6-8 **Small-*N* Designs and Hypothesis Investigation**


Hypothesis suggested by large-N study, general observation, or other sources can be investigated in detail by small-N research.
Small-N study: one or few cases; many observations or measurements
Hypothesis suggested by small-N study can be tested in another design involving more cases (e.g., survey of 100 organizations' leadership styles).

COMPARATIVE STUDY. This kind of research may involve more than one case; such studies are often called *comparative case studies.* A comparative or multiple case study is more likely to have explanatory power than is a single case study because it provides the opportunity for replication; that is, it enables a researcher to test a single theory more than once. For some cases, similar results will be predicted; for others, different results will be predicted.[32] Multiple cases should not be thought of as a "sample," because cases are not chosen using a statistical procedure to form a "representative" sample from which the frequency of a particular phenomenon will be calculated and inferences about a larger population drawn. Rather, cases are chosen for the presence or absence of factors that a political theory has indicated are important.

As an example of the logic and layout of a comparative study and its potential utility in causal analysis, suppose a political scientist wanted to know why socialism never emerged as a major political force in the United States, especially compared to European nations like Great Britain—a situation that has intrigued innumerable

[32] Yin, *Case Study Research: Design and Methods,* 53.

scholars for the past one hundred years.[33] The most commonly cited "causes" of the failure of socialism to take root in America include, among other things, "its relatively high levels of social equalitarianism, [enormous] economic productivity, and social mobility (particularly into elite strata), alongside the strength of religion, the weakness of the central state, the earlier timing of electoral democracy, ethnic and racial diversity, and . . . the absence of fixed social classes."[34] Imagine the researcher trying to sort out these possibilities by comparing France and the United States.

One strategy is to apply the *method of difference*, introduced by the English philosopher John Stuart Mill: "If an instance in which the phenomenon under investigation occurs, and an instance in which it does not occur, have every circumstance in common save one, that one occurring only in the former: the circumstance in which alone the two instances differ, is the effect, or cause, or a necessary part of the cause, of the phenomenon."[35] For example, our investigator would first identify a country in which the condition (socialism) is present in and one which it is not, and he or she would then look for similarities and differences in these antecedents. As the hypothetical data in table 6-11 suggest, it might be the case that in the nineteenth century, the United States and France shared similar experiences such as extensive industrialization and urbanization and that in both countries citizens spoke common languages (French and English), but they *differed* in that the French had a fixed and rigid social class system (including a landed aristocracy) whereas America did not. Since the two countries have parallel backgrounds except for their class structures, we may infer that this difference rather than the other factors explains why socialism has not had much influence on American politics.

Needless to say, this comparison is woefully inadequate and simplistic. In fact, no real analysis would take exactly this form. Instead, the table is a "reconstruction" of the logic of comparison using this method. (Mill, by the way, introduced several other comparative methods, but knowing them is not essential for understanding the gist of comparative analysis.[36]) If you were to attempt research of this sort, you would have to consider many more factors and make difficult decisions about when an antecedent is or is not present. But the method of difference and similar designs

[33] See, among countless other sources, Werner Sombart, *Why Is There No Socialism in the United States?* (White Plains, N.Y.: International Arts and Sciences Press, 1976), first published in German in 1906; and Seymour Martin Lipset, *The First New Nation: The United States in Historical and Comparative Perspective* (New York: Basic Books, 1963).

[34] Seymour Martin Lipset and Gary Marks, *It Didn't Happen Here: Why Socialism Failed in the United States* (New York: W. W. Norton, 2000), 16.

[35] John Stuart Mill, *A System of Logic, Ratiocinative and Inductive: Being a Connected View of the Principles of Evidence and the Methods of Scientific Investigation* (New York: Harper, 1850), 225. Available at http://www.archive.org/details/systemoflogicrat01milliala

[36] See Merrilee H. Salmon, *Introduction to Logic and Critical Thinking*, 2nd ed. (San Diego, Calif.: Harcourt Brace Jovanovich, 1989), 109–15.

TABLE 6-11 **Mill's Method of Difference**

Case (country)	Socialist movements?	Condition or Effect Antecedent 1 (industrialized)	Antecedent 2 (urbanized)	Antecedent 3 (common language)	Antecedent 4 (historically strong and fixed social classes)
United States	no	yes	yes	yes	no
France	yes	yes	yes	yes	yes

underlies a great deal of political research.[37] By the way, note that one can interpret this approach as a natural experiment.

Despite case studies' potential to make important contributions to our understanding of political phenomena, there are some concerns about the knowledge they generate.[38] One potential problem is the "lack of rigor" in presenting evidence and the possibility for bias in using it. Typically, researchers sift through enormous quantities of detailed information about their cases. But how does one know all the important possible antecedents have been identified? Has something significant been omitted? Or the researcher may be the only one to have recorded certain behavior or phenomena. Still, the potential for bias of this sort is not limited to case studies.

Another frequently raised criticism of case studies is the problem of generalization. One response to this criticism is to use multiple case studies. In fact, as Yin pointed out, the same criticism can be leveled against a single experiment: scientific knowledge is usually based on multiple experiments rather than on a single experiment.[39] Yet people do not say that performing a single experiment is not worthwhile. Furthermore, Yin stated that

> Case studies, like experiments, are generalizable to theoretical propositions and not to populations or universes. In this sense, the case study, like the experiment, does not represent a "sample," and the investigator's goal is to expand and generalize theories (analytic generalization) and not to enumerate frequencies (statistical generalization).[40]

[37] Skocpol, *States and Social Revolutions*, is an excellent example of seminal research that explicitly uses Mill's method.

[38] Yin, *Case Study Research: Design and Methods*, 21–22.

[39] Ibid., 21.

[40] Ibid., 23.

A third potential drawback of case studies is that they may require long and arduous efforts to describe and report the results owing to the need to present adequate documentation. (Think about the complexity of untangling the differences between French and American societies.) This criticism may stem from confusing the case study with particular methods of data collection, such as participant observation (discussed in chapter 8), which often requires a long period of data collection.[41] However, case studies should not be ruled out as an appropriate research design due to this historic association.

Finally, in spite of the enthusiasm for case studies, considerable debate remains about just how strong causal inferences can be in these designs. Consider table 6-11 again.[42] If our hypothetical study had discovered that France and the United States had the same values on *all* the independent variables, we would conclude that none of the hypotheses in this instance holds. And, as we stressed in chapter 2, that might be an important conclusion, given that falsification of propositions is one of the goals of science. But instead the conclusion seems to be that having a deep-seated social class system is at least a necessary condition for the emergence of socialism. Yet the result is hardly definitive. If, for example, the "real" cause of the dependent variable is not identified and explicitly included in the analysis, this design cannot detect it. Or it is plausible that the nonexplanatory variables (e.g., industrialization, common language) "interact" or have a simultaneous joint effect on the dependent variable. In a design of the sort illustrated in table 6-11, it is impossible to know.[43] Furthermore, this argument assumes causation is deterministic: once *X* appears, *Y always* follows. Yet many, if not most, scholars regard probabilistic causation—if *X* appears, then *Y probably* follows—as a more realistic description of the way the world works. Is it possible that deep socioeconomic cleavages do not always produce socialist movements? The method of difference provides no foolproof answer.[44]

Still, in many circumstances the case study design can be an informative and appropriate research design. The design permits a deeper understanding of causal

[41] Ibid.

[42] The literature discussing the pros and cons of small-*N* research, especially its applicability to causal inference, includes, among many others, Gary King, Robert Keohane, and Sidney Verba, *Designing Social Inquiry: Scientific Inference in Qualitative Research* (Princeton, N.J.: Princeton University Press, 1994); James Mahoney, "Strategies of Causal Analysis in Small-*N* Analysis," *Sociological Methods and Research* 28, no. 4 (2000): 387–424; and Stanley Lieberson, "Small *N*'s and Big Conclusions: An Examination of the Reasoning in Comparative Studies Based on a Small Number of Cases," *Social Forces* 70, no. 2 (1991): 307–20, available at http://www.wjh.harvard.edu/soc/faculty/lieberson/Small_Ns_and_Big_Conclusions.pdf

[43] Lieberson, "Small *N*'s and Big Conclusions," 312–13.

[44] For an extended discussion, see King, Keohane, and Verba, *Designing Social Inquiry: Scientific Inference in Qualitative Research;* and Douglas Dion, "Evidence and Inference in the Comparative Case Study," *Comparative Politics* 30, no. 2 (1998): 127–45. They present a more optimistic picture of the possibilities of causal inference in small-*N* research.

processes, the explication of general explanatory theory, and the development of hypotheses regarding difficult-to-observe phenomena. Much of our understanding of politics and political processes comes from case studies of individual presidents, senators, representatives, mayors, judges, statutes, campaigns, treaties, policy initiatives, and wars. The case study design should be viewed as complementary to, rather than inconsistent with, other experimental and nonexperimental designs.

FOCUS GROUP. A focus group consists of a small number of individuals (about ten to twenty, say) who meet in a single location and discuss with a leader a topic or research stimulus such as a proposed campaign brochure. A focus group can superficially resemble an experiment, but no effort is usually made to assign participants randomly to treatment and control groups or to systematically introduce an experimental variation. The deliberations may or may not be (surreptitiously) recorded or observed by others on the research team. This approach lends itself nicely to market research, but for the reasons just mentioned is seldom used to make causal inferences.

Focus groups have become somewhat controversial in politics because, critics assert, the results often encourage candidates, groups, and parties to take "safe" or noncontroversial positions on issues. Some seemingly daring policy proposals, for example, have been thoroughly researched in focus groups to ease the minds of political consultants that their candidates stood very little risk by adopting them. Yet these small-group discussions can be used to create hypotheses that can then be tested in larger surveys. Suppose you want to conduct a poll on your campus about physician-assisted suicide. You might begin with a focus-group discussion to see generally what students think about the issue. The verbal reports might then assist you in developing some specific items to place in a questionnaire.

Cross-Sectional Designs: Surveys and Aggregate Analysis

Perhaps the most common nonexperimental research design is cross-sectional analysis. In a **cross-sectional design**, measurements of the independent and dependent variables are taken at approximately the same time,[45] and the researcher does not control or manipulate the independent variable, the assignment of subjects to treatment or control groups, or the conditions under which the independent variable is experienced. If the units of analysis are individuals, the study is often called a *survey* or poll; if the subjects are geographical entities, such as states or nations or other groupings of units, the term *aggregate analysis* is frequently applied. In either

[45] Although the measurements may be taken over a period of days or even weeks, cross-sectional analysis treats them as though they were obtained simultaneously.

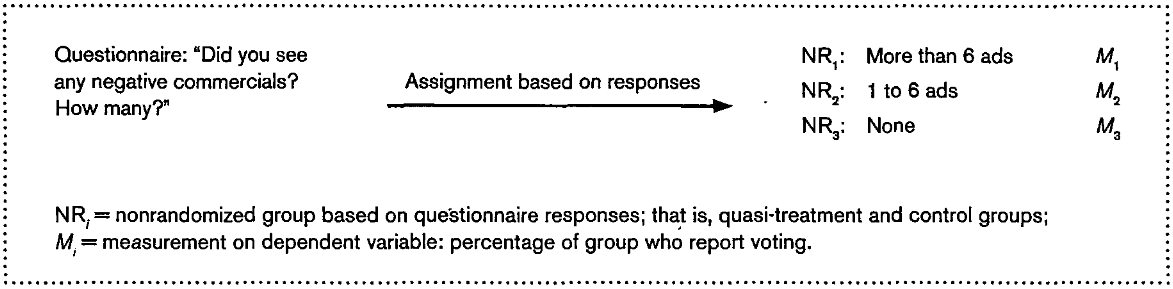
situation, the units are simply measured or observed and the data recorded. In surveys, the respondents themselves report their exposure to various factors. In aggregate analysis, the investigator only observes which units have what values on the variables. The measurements are used to construct, with the help of statistical methods, posttreatment quasi-experimental and quasi-control groups that have naturally occurred, and the measurements of the dependent variable are used to assess the differences between these groups. Data analysis, rather than physical manipulation of variables, is the basis for making causal inferences.

Although this approach makes it far more difficult to measure the causal effects that can be attributed to the presence or introduction of independent variables (treatments), it has the virtues of allowing observation of phenomena in more natural, realistic settings; increasing the size and representativeness of the populations studied; and allowing the testing of hypotheses that do not lend themselves easily to experimental treatment. In short, cross-sectional designs improve external validity at the expense of internal validity.

The example presented at the beginning of the chapter illustrates a particularly simple cross-sectional study. Recall that we tried to assess the effects of negative campaigning on the likelihood of voting by interviewing (that is, surveying) a sample of citizens and then dividing the respondents into different categories according to *their* answers to questions of this sort: "Did you happen to see or read any campaign ads? How many? How many seemed to attack the opponent?" We could then sort respondents by their self-reported level of exposure to negativity. Notice that since there is no random assignment, only self-reports, we do not control who is in each group by forcing people to view differing levels of negative advertising. The groups are simply observed. (Figure 6-9 shows a "reconstructed" layout.) If the groups differed by their rate of voting, we would have a relationship but would not demonstrate cause and effect. Suppose, for instance, we find that M_1 (the percentage of those who viewed six or more negative ads who also reported voting) is less than M_2, which in turn is less than M_3?

Because of our research design and our inability to ensure that those with less and those with more exposure were alike in every other way, we could not necessarily conclude that campaign tone determines the propensity to vote. And note that this is true no matter how large our sample is. With a survey design, then, we typically have to employ data analysis techniques to control for potential confounders that may affect both the independent and dependent variables. If we wanted to control for these factors, we would have to include appropriate questions in the survey and then use statistics to hold them constant. Suppose we thought that education independently affected the propensities to watch a lot of television and to not vote. In a survey, we include a question about the level of the respondents' schooling, as indicated in figure 6-10. Here we have formed six, nonrandomized groups and can

FIGURE 6-9 **Logic of Survey Design**



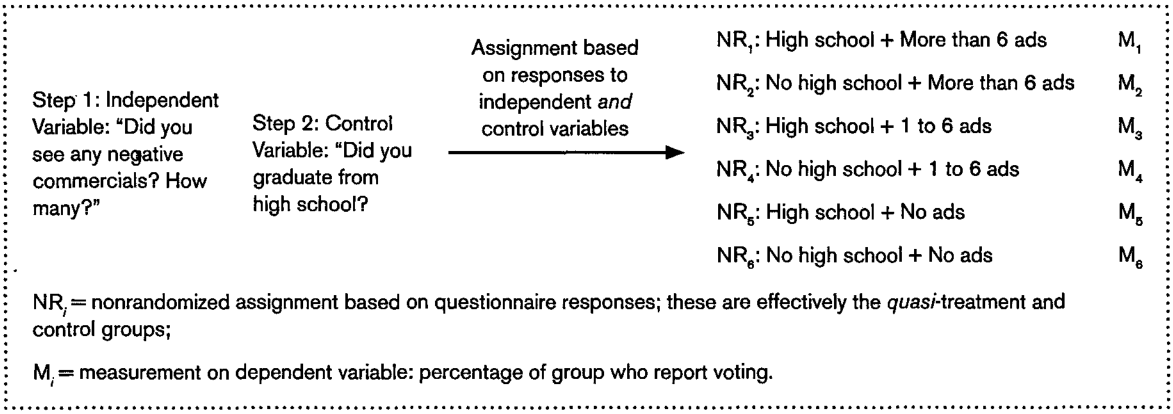
Questionnaire: "Did you see any negative commercials? How many?"

Assignment based on responses →

NR_1:	More than 6 ads	M_1
NR_2:	1 to 6 ads	M_2
NR_3:	None	M_3

NR_i = nonrandomized group based on questionnaire responses; that is, quasi-treatment and control groups; M_i = measurement on dependent variable: percentage of group who report voting.

FIGURE 6-10 **Design with Control Variable**



Step 1: Independent Variable: "Did you see any negative commercials? How many?"

Step 2: Control Variable: "Did you graduate from high school?"

Assignment based on responses to independent *and* control variables →

NR_1: High school + More than 6 ads	M_1
NR_2: No high school + More than 6 ads	M_2
NR_3: High school + 1 to 6 ads	M_3
NR_4: No high school + 1 to 6 ads	M_4
NR_5: High school + No ads	M_5
NR_6: No high school + No ads	M_6

NR_i = nonrandomized assignment based on questionnaire responses; these are effectively the *quasi*-treatment and control groups;

M_i = measurement on dependent variable: percentage of group who report voting.

compare their participation rates, *M*, as before. Presumably, if education is creating the (spurious) relationship between voting and viewing habits, the *M*s would all be about the same except for sampling and measurement error. If, however, advertising does affect motivations even after controlling for education, the average measurements in the groups would vary (e.g., under the hypothesis being considered, M_1 and M_2 would be less than, say, M_3 and M_4).

In essence, the limitations of the cross-sectional design—that is, lack of control over exposure to the independent variable and inability to form pure experimental and control groups—force us to rely on data analysis techniques to isolate the impact of the independent variables of interest. This process requires researchers to make their comparison groups equivalent by holding relevant extraneous factors constant and then observing the relationship between independent and dependent variables, a procedure described more fully in chapter 14. Yet holding these factors constant is problematic, since it is very difficult to be sure that all relevant variables have been explicitly identified and measured. It is important to stress that if a causal variable is not recognized and brought into the analysis, its effects are nonetheless still operative, even though we may not be aware of them.

Longitudinal (Time Series) Designs

Longitudinal or **time series designs** are characterized by the availability of measures of variables at different points in time. As with the other designs, the researcher does not control the introduction of the independent variable(s) and must rely on data collected by others to measure the dependent variable rather than personally conducting the measurements. On the other hand, time series designs have two distinct advantages: (1) change in the level of variables or conditions can be measured and modeled, and (2) it is sometimes easier to decide time order or which comes first, X or Y.

Additional benefits of longitudinal studies include the fact that they can in principle estimate three kinds of effects: age, period (history), and cohort. Age effects can be considered a direct measure of (chronological) time and be assessed like other variables. As in cross-sectional work, an investigator may be interested in the effect of age on political predispositions or ideology. (It is commonly asserted that as people age, they become more politically conservative.) But in addition, in longitudinal analysis a period (interval of time) may be thought of as an indicator of history during a period, and the consequences on individuals are **period effects**. It is the "history" that occurs during the period, not chronological age, that matters. During the late 1960s and early 1970s, for example, events such as Watergate and the Vietnam War adversely affected many citizens' trust in government, whether they were young or old. When that era passed, its effects on newer generations dissipated. So those who lived through those stormy times might have very different beliefs and opinions than do younger people.

Another way of interpreting period effects is to consider cohorts. A **cohort** is defined as a group of people who all experience a significant event in roughly the same time. A birth cohort, for instance, consists of those born in a given year or period; an "event" cohort is those who shared a common experience, such as their first entry into the labor force at a particular time. It is often hypothesized that individuals in one cohort will, because of their shared background, behave differently than individuals in a different cohort. To take one example, people born in the years immediately after World War II (the baby boomers) may have different political attitudes and affiliations than those who were born in the 1980s.

Note that cohort, period, and age effects are inescapably related because "cohort (year of birth) = period (year of event) + age (years since birth)."[46] There are, in short, a number of ways of understanding longitudinal research; the choice depends on the analyst's interests.

TREND ANALYSIS. Former Congressman Lee Hamilton noted, "There's a funny thing going on in our national politics right now: Everyone deplores polarization, but it just keeps getting worse."[47]

Indeed, a quick survey of the political landscape *seems* to indicate a deep and growing divide between conservatives and liberals and between Republicans and Democrats. Its existence is conventional wisdom. Surprisingly, perhaps, some academic research finds that claims of a wide and widening chasm between Americans may be overdrawn.[48] To know if Americans are becoming more polarized—more sharply divided between strong liberals and strong conservatives, with relatively few in the middle—we must answer three questions:

1. Exactly who is polarized?
2. What does *polarized* mean?
3. Has the division been growing?

We can answer the first two questions quickly. Political scientists distinguish between polarization among elites and among the general public. There is evidence that among leaders in Washington, D.C., especially, the two parties have become more partisan and the ideological divide between them is as broad as it has ever been. But the same may not be true for average citizens. How do we measure the distance between the ideological groups? Again, political scientists turn to questionnaires and survey data. The General Social Survey, a collaborative research project housed at the National Opinion Research Center (NORC) at the University of Chicago, has been conducting national surveys on a yearly basis for

[46] See Scott Menard, *Longitudinal Research,* A Sage University Paper: Quantitative Applications in the Social Sciences no. 07–076 (Newbury Park, Calif.: Sage, 1991), 7.

[47] Lee H. Hamilton, "The Changes Necessary to Make American Politics Less Polarized," *Deseret News,* December 6, 2010. Available at http://www.deseretnews.com/article/700088722/The-changes-necessary-to-make-American-politics-less-polarized.html

[48] See, among others, Alan I. Abramowitz, *The Disappearing Center: Engaged Citizens, Polarization, and American Democracy* (New Haven, Conn.: Yale University Press, 2010); Geoffrey C. Layman and Thomas M. Carsey, "Party Polarization and 'Conflict Extension' in the American Electorate," *American Journal of Political Science* 46, no. 4 (2002): 786–802; Morris P. Fiorina and Samuel J. Abrams, "Political Polarization in the American Public," *Annual Review of Political Science* 11 (2008): 563–88, available at http://www.sociology.uiowa.edu/nsfworkshop/JournalArticleResources/Fiorina_Abrams_Political_Polariza tion_2008.pdf

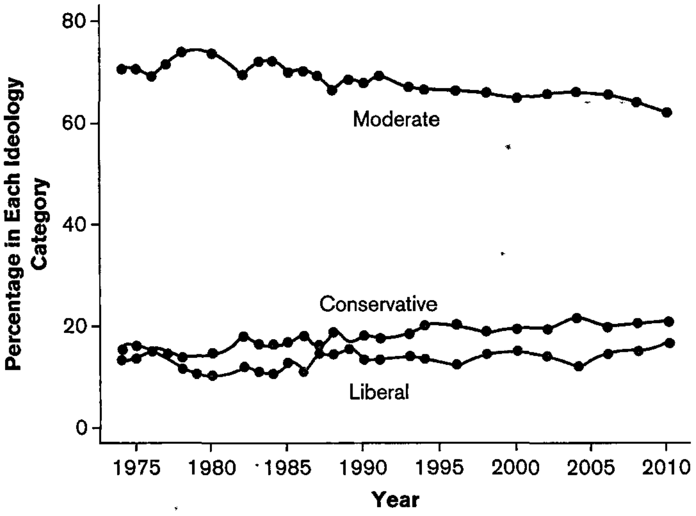
more than two decades. Many of the questionnaires across the years contained this question:

> We hear a lot of talk these days about liberals and conservatives. I'm going to show you a seven-point scale on which the political views that people might hold are arranged from extremely liberal—point 1—to extremely conservative—point 7. Where would you place yourself on this scale?[49]

For purposes of our study, we categorized respondents into four categories: strong liberal, liberal, moderate (including "slightly" liberal and "slightly" conservative individuals), conservative, and strong conservative.

To address the third question, we turn to **trend analysis**. As the term implies, the analysis of a trend starts with measurements or observations on a dependent variable of interest taken at different times (usually twenty or more) and attempts to determine whether and why the level of the variable is changing. A simple approach for numeric data is to plot some appropriate summary measure of the dependent variable at different times. Figure 6-11 shows the percentage of respondents in each

FIGURE 6-11 **Political Ideology Self-Identification, 1974–2010**



Source: James A. Davis, Tom W. Smith, and Peter V. Marsden, *General Social Surveys, 1972–2010* (Chicago: National Opinion Research Center, 2011).

Note: Ideology grouped in three categories.

[49] GSS 1972–2008 Cumulative Dataset, http://www.norc.org/GSS+Website/

ideology category for twenty-five years from 1974 to 2010. If there is increased polarization, we should be able to spot it in the graph. And we see immediately that all the lines are more or less flat over the period; the percentage of moderates has decreased slightly, while the percentages of liberals and conservatives have risen slightly.

By itself, a graph of one variable over time cannot tell us *why* a variable is trending up or down or even moving randomly. For that analysis, we need slightly more advanced statistical tools and more data. For the latter purpose, an investigator needs to introduce additional variables and measure them over time. This type of analysis takes (roughly speaking) this form:

$$Y_t = f\,(Y_{t-1} + Y_{t-2} \ldots + Xs_t + Xs_{t-1} + Xs_{t-2} \ldots),$$

where the *Y*s and *X*s are measures of the dependent (*Y*) and independent (*X*) variables at the current (latest) time (*t*) and at previous times (*t*-1, *t*-2, . . . etc.) and *f* means "is a function of" or "is produced by."[50] When data are measured at many different points, as illustrated in figure 6-11 above, statistical procedures called *time series analysis* are often employed. Note also that, although the previous examples pertain to changing proportions in samples of individuals, trends in aggregate variables (e.g., crime or poverty rates in urban areas) can also be investigated.

Conclusion

In this chapter, we have discussed why choosing a research design is an important step in the research process. A design enables the researcher to achieve his or her research objectives and can lead to valid, informative conclusions. We presented two basic types of research designs—experimental and nonexperimental—along with a couple of alternative approaches. We discussed their advantages and disadvantages. Experimental designs—which allow the researcher to exercise control over the independent variable, the units of analysis, and their environment—are often preferred over nonexperimental designs because they enable the researcher to establish sounder causal explanations. Therefore, experimental designs are generally stronger in internal validity than nonexperimental ones. However, it may not always be possible or appropriate to use an experimental design. Thus, nonexperimental observation may also be used to test hypotheses in a meaningful fashion and often in a way that increases the external validity of the results. In these instances, causal assertions rest on weaker grounds and frequently have to be approximated by statistical means (see chapter 13). Yet the basic objectives of research designs, whether experimental or nonexperimental, are the same.

[50] In practice, relationships of this sort are thought of as probabilistic, not deterministic, so a random error term would be added.

TERMS INTRODUCED

Case study design. A comprehensive and in-depth study of a single case or several cases. A nonexperimental design in which the investigator has little control over events.

Classical randomized experimental design. An experiment with the random assignment of subjects to experimental and control groups with a pretest and posttest for both groups.

Cohort. A group of people who all experience a significant event in roughly the same time frame.

Control group. A group of subjects that does not receive the experimental treatment or test stimulus.

Correlation. A statement that the values or states of one thing systematically vary with the values or state of another; an association between two variables.

Cross-sectional design. A research design in which measurements of independent and dependent variables are taken at the same time; naturally occurring differences in the independent variable are used to create quasi-experimental and quasi-control groups; extraneous factors are controlled for by statistical means.

Demand characteristics. Aspects of the research situation that cause participants to guess the purpose or rationale of the study and adjust their behavior or opinions accordingly.

Experiment. Research using a research design in which the researcher controls exposure to the test factor or independent variable, the assignment of subjects to groups, and the measurement of responses.

Experimental effect. Effect, usually measured numerically, of the experimental variable on the dependent variable.

Experimental group. A group of subjects that receives the experimental treatment or test stimulus.

Experimental mortality. A differential loss of subjects from experimental and control groups that affects the equivalency of groups; threat to internal validity.

External validity. The ability to generalize from one set of research findings to other situations.

Field experiment. Experimental designs applied in a natural setting.

Internal validity. The ability to show that manipulation or variation of the independent variable actually causes the dependent variable to change.

Intervention analysis. A nonexperimental time series design in which measurements of a dependent variable are taken both before and after the "introduction" of an independent variable.

Multiple-group design. Experimental design with more than one control and experimental group.

Natural experiment. A study in which comparisons are made among "naturally" occurring groups on variables that cannot be controlled by the investigator.

Period effect. An indicator or measure of history effects on a dependent variable during a specified time.

Policy evaluation. Objective analysis of economic, political, cultural, or social effects of public policies.

Posttest design. Research design in which the dependent variable is measured after, but not before, manipulation of the independent variable.

Pretest. Measurement of the dependent variable prior to the administration of the experimental treatment or manipulation of the independent variable.

Quasi-experimental design. A research design that includes treatment and control groups to which individuals are not assigned randomly.

Randomization. The random assignment of subjects to experimental and control groups.

Repeated-measurement design. A plan that calls for making more than one measure or observation on a dependent variable at different times over the course of the study.

Research design. A plan specifying how the researcher intends to fulfill the goals of the study; a logical plan for testing hypotheses.

Selection bias. Bias due to the assignment of subjects to experimental and control groups according to some criterion and not randomly; threat to internal validity.

Small-*N* design. A research design in which the researcher examines one or a few cases of a phenomenon in considerable detail.

Test stimulus or test factor. The independent variable introduced and controlled by an investigator in order to assess its effects on a response or dependent variable.

Time series design. A research design (sometimes called a longitudinal design) featuring multiple measurements of the dependent variable before and after experimental treatment.

Trend analysis. Research design that measures a dependent variable at different times and attempts to determine whether the level of the variable is changing and, if it is, why.

SUGGESTED READINGS

Campbell, Donald T., and Julian C. Stanley. *Experimental and Quasi-Experimental Designs for Research.* Chicago: Rand McNally, 1966.

Cook, Thomas D., and Donald T. Campbell. *Quasi-Experimentation: Design and Analysis Issues for Field Settings.* New York: Houghton Mifflin, 1979.

Downs, Anthony. *An Economic Theory of Democracy.* New York: Harper and Row, 1957.

Gerring, John. "What Is a Case Study and What Is It Good For?" *American Political Science Review* 98, no. 2 (2004): 341–54.

Hakim, Catherine. *Research Design: Strategies and Choices in the Design of Social Research.* Contemporary Social Research Series no. 13. London, UK: Allen and Unwin, 1987.

King, Gary, Robert O. Keohane, and Sidney Verba. *Designing Social Inquiry: Scientific Inference in Qualitative Research.* Princeton, N.J.: Princeton University Press, 1994.

Laver, Michael. *Private Desires, Political Action: An Invitation to the Politics of Rational Choice.* Thousand Oaks, Calif.: Sage, 1997.

Menard, Scott. *Longitudinal Research.* A Sage University Paper: Quantitative Applications in the Social Sciences no. 07–076. Newbury Park, Calif.: Sage, 1991.

Sambanis, Nicholas. "Using Case Studies to Expand Economic Models of Civil War." *Perspectives on Politics* 2, no. 2 (2004): 259–79. Available at http://www.apsanet.org/imgtest/sambanis PoP (june 04).pdf

Sekhon, Jasjeet S. "Quality Meets Quantity: Case Studies, Conditional Probability, and Counterfactuals." *Perspectives on Politics* 2, no. 2 (2004): 281–93. Available at http://sekhon.berkeley.edu/papers/QualityQuantity.pdf

Spector, Paul E. *Research Designs.* A Sage University Paper: Quantitative Applications in the Social Sciences no. 07–023. Beverly Hills, Calif.: Sage, 1981.

Vandaele, Walter. *Applied Time Series and Box-Jenkins Models.* New York: Academic Press, 1983.

Yin, Robert K. *Case Study Research: Design and Methods.* Rev. ed. Applied Social Research Methods Series, vol. 5. Newbury Park, Calif.: Sage, 1989.

CHAPTER 7

Sampling

CHAPTER OBJECTIVES

7.1 Describe how sampling works.

7.2 Identify five different types of samples.

7.3 Explain what can be learned from a population sample and its limitations.

IN A STUDY OF WHY PEOPLE do and do not participate in surveys, researchers at the Pew Research Center found that both willing and reluctant participants expressed skepticism about polling validity: "many in each group (65% and 68%, respectively) doubted that a random sample of 1,500 people can 'accurately reflect the views' of the American public."[1] Pollingreport.com ("an independent, nonpartisan resource on trends in American public opinion") finds one source of doubt is the reliability and validity of claims made on the basis of a sample of a much larger population:

> How can a sample of only 800 or 1,200 truly reflect the opinions of [300+] million Americans within a few percentage points?[2]

In chapter 1, for example, Kriner and Shen's study of public support for military engagements and to what extent support was affected by the number of expected casualties as well as whether those sacrifices were borne by all segments of society. Of course, putting these questions to every citizen

[1] Pew Research Center for the People and the Press, "Possible Consequences of Non-Response for Pre-Election Surveys: Race and Reluctant Respondents," May 16, 1998. Available at http://people-press.org/1998/05/16/possible-consequences-of-non-response-for-pre-election-surveys/

[2] National Council on Public Polls, "Answers to Questions We Often Hear from the Public." Accessed January 7, 2015. Available at http://www.pollingreport.com/ncpp.htm

HELPFUL HINTS

"Population"

Do not be confused by the term *population,* which, as the text indicates, means simply a collection of things. We could define a population as the people living in New Castle, Delaware. But a population could also consist of a set of geographical areas, such as the voting districts in New Castle County. In the first case, the units of analysis are individuals; in the second case, they are aggregates of individuals.

Check out more Helpful Hints at **edge.sagepub.com/johnson8e**

would be impractical. So most researchers collect information on a much smaller set of individuals. As we just noted, however, that strategy raises another issue: If an investigation of public opinion rests on 100 or even 1,000 observations, can it really say anything about the millions of Americans who comprise the general public? Can it, in other words, lead to reliable and valid conclusions?

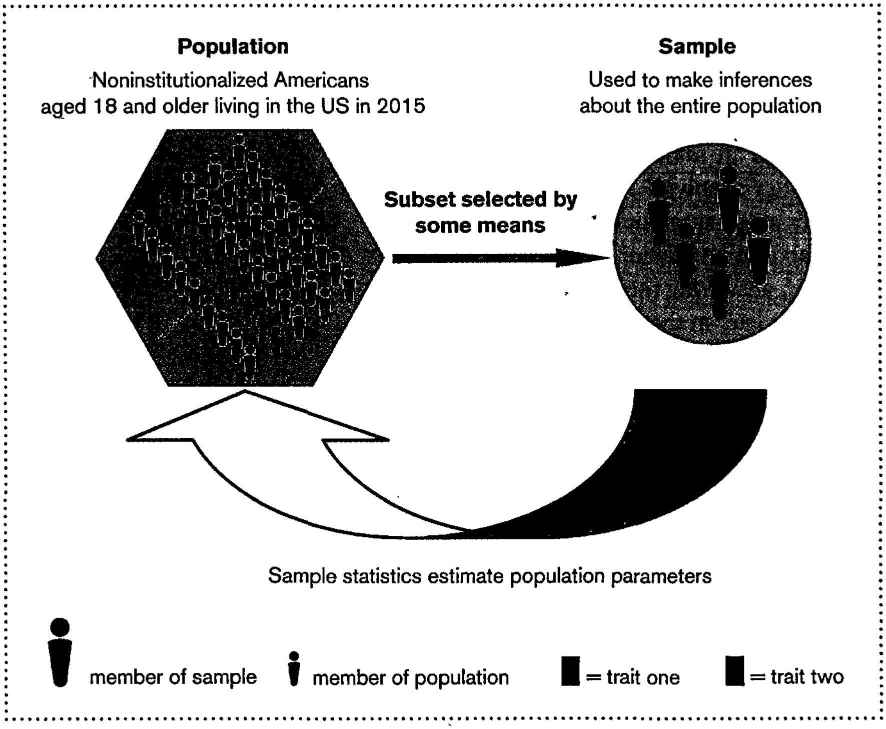
Our task in this chapter is to provide an answer to two general questions. First, exactly what are samples, and how are they collected? Second, what kind of information do they supply? Do they really provide precise measures of opinions, or do they just offer rough approximations? That is, how much confidence can we place in statements about a population given observations derived from a very few of its members? We begin answering these questions with a description of sampling techniques and reserve for later in the chapter a discussion of inferences based on samples.

The Basics of Sampling

The fundamentals are quite simple, at least in theory (see figure 7-1). Suppose we want to assess Americans' level of support for a military action. At the outset we need to clarify what we mean by *Americans.* More formally, we need to define or specify an appropriate population. In the figure the population is defined to be all adult (aged eighteen and older) citizens not residing in institutional settings (for example, prisons, hospitals) in the United States in 2015. A **population** is any well-defined set of units of analysis. It does not necessarily refer to people. A population might be all the adults living in a geographical area, such as a country or state, or working in an organization. But it could equally well be a set of counties, corporations, government agencies, events, magazine articles, or years. What is important is that the population be carefully and fully defined and that it be relevant to the research question.[3] The polygon in figure 7-1 represents the population of adult American citizens. Since there are millions and millions of citizens, the diagram only symbolizes this huge number. In this hypothetical analysis our claims refer to these people, not to Germans or Mexicans or children or any other group.

Since it is impossible to interview everyone, a more practical approach is to select just a "few" members of the population for further investigation. This is where sampling comes in. A **sample** is any subset of units collected in some manner from a population. (In the figure, the sample consists of just five out of millions of people.) The sample size and *how* its members are chosen determine the quality (that is, the accuracy and reliability) of inferences about the whole population. The important things to clarify are the method of selection and the number of observations to be drawn.

FIGURE 7-1 **Population and Sample**


Population
Noninstitutionalized Americans aged 18 and older living in the US in 2015
Sample
Used to make inferences about the entire population
Subset selected by some means
Sample statistics estimate population parameters
member of sample
member of population
■ = trait one
■ = trait two

[3] A related concern is the size of the population. In fact, no population of real "things" has an infinite number of members, but we nevertheless treat populations as if they were infinite for most purposes.

Once a sample has been gathered, features or characteristics of interest can be examined and measured. The attributes of most interest in empirical research are numerical or quantitative indicators such as percentages or averages. These measures—or **sample statistics**, as they are known—are used to approximate the corresponding population values, or parameters. That's the idea behind the arrow: we use sample statistics to estimate population characteristics (parameters). It may be intuitively obvious that the sample statistics will not exactly equal the corresponding population values. But, as we demonstrate in this chapter, if we follow suitable procedures, they will be reasonably close.

Population or Sample?

A researcher's decision whether to collect data for a population or for a sample is usually made on practical grounds. If time, money, and other costs were not considerations, it would almost always be better to collect data for a population, because we would then be sure that the observed cases accurately reflected the population characteristics of interest. However, in many if not most instances it is simply not possible or feasible to study an entire population. Since research is costly and time-consuming, researchers must weigh the advantages and disadvantages of using a population or a sample. The advantages of taking a sample are often savings in time and money. The disadvantage is that information based on a sample is usually less accurate or more subject to error than is information collected from a population.

HELPFUL HINTS

Is a Sample Always Less Accurate?

In the late 1990s Congress and President Bill Clinton debated the merits of using sampling instead of trying to interview the entire population when conducting the 2000 Census. Clinton and the Census Bureau argued that the methods used to tally everyone leads to so many errors that many groups are undercounted—in particular, undocumented aliens and inhabitants of inner cities and rural areas. It would be more accurate, they maintained, to draw careful samples of target populations and conduct quality interviews and measurements. But members of Congress (mainly Republicans) argued that the Constitution requires a complete enumeration of the people. (Politics formed the context of this dispute; after all, innumerable government grants as well as seats in the House of Representatives are awarded according to population size, and undercounting is thought to be more of a problem in traditionally Democratic areas.)

Check out more Helpful Hints at **edge.sagepub.com/johnson8e**

Often, researchers do not have the option of studying a complete population. Consider, as an example, a political scientist who wants to test some hypotheses regarding the content of televised political campaign commercials. The project requires an examination of the content of numerous commercials, which is the unit of analysis. From the standpoint of accuracy, it would be preferable to have data on the total population of televised commercials (in other words, to have available for measurement every campaign commercial that ever aired). But undertaking this type of analysis is simply impossible because no such data bank exists anywhere, nor does anyone even know how many such commercials have been televised across the country since they first appeared. Consequently, the researcher will have to rely on a sample of readily available commercials to test the hypotheses—a decision that is practical, necessary, and less costly, but perhaps subject to error.[4]

Fundamental Concepts

As noted in the previous section, a sample is simply a subset of a larger population, just as a sample of blood is a subset of all the blood in your body at one moment in time. If the sample is selected properly, the information it yields may be used to make inferences about the whole population. Since sampling is always used in public opinion surveys, it is often thought of in connection with that activity. Sampling arises whenever a researcher takes measurements on a subset of a population covered by the hypothesis being investigated. Whatever empirical findings emerge from a sample from a specified population, however, will apply to that and only that

Notation

Population parameters are typically denoted by capital Roman or Greek letters. A proportion, such as the proportion of Americans who support President Obama at a particular time, typically is designated *P* or π (the Greek letter *pi*). The purpose of sampling is to collect data that provide an accurate inference about a population parameter. An **estimator**, then, is a sample statistic based on sample observations that estimates the numerical value of a population characteristic, or parameter. A specific estimator of a population characteristic or attribute calculated from sample data is called a sample statistic. Like population parameters, these are typically denoted by symbols or letters. Frequently, we use a hat (^) over a character to denote a sample statistic; in some situations, a lowercase letter is used—for example, a lowercase *p* for a sample proportion. Sometimes, though, another symbol is used. The population mean (average) is almost always symbolized by μ (the lowercase Greek *mu*). But in this case nearly everyone lets $\bar{Y}$, not $\hat{\mu}$, stand for the sample mean.

[4] Richard A. Joslyn, *Mass Media and Elections* (Reading, Mass.: Addison-Wesley, 1984).

population. It would be a mistake, for instance, to sample campaign speeches from the last four presidential elections and then generalize to all American presidential rhetoric. By contrast, a sample drawn from a population of campaign speeches given by all presidents could be used to generalize to that population of speeches.

Before proceeding further, we should note that what usually matters most is that samples are obtained according to well-established rules. To understand why, we need to review some terms commonly used in discussions of sampling.[5]

An **element** (frequently called a unit of analysis) is a single occurrence, realization, or instance of the objects or entities being studied. Elements in political science research are often individuals, but they also can be states, cities, agencies, countries, campaign advertisements, political speeches, wars, social or professional organizations, crimes, or legislatures, just to name a few.

As noted previously, a population is a collection of elements defined according to a researcher's theoretical interest. Sometimes this is referred to as the theoretical population. It may, for example, consist of all campaign speeches given by major candidates for president in the last four presidential elections. Or it may be all international armed conflicts that have occurred in the past two hundred years. The key is to be clear and specific. You may refer to presidential campaign speeches as the focus of your research, but at some point you should make clear which speeches in what time periods constitute the population.

For reasons that we discuss shortly, a population may be stratified—that is, subdivided or broken up into groups of similar elements—before a sample is drawn. Each **stratum** or layer is a subgroup of a population that shares one or more characteristics. For example, we might divide the population of campaign speeches in the last four presidential elections into four strata, each stratum containing speeches from one of the four elections. In a study of students' attitudes, particularly at a university, the student body may be stratified by academic class, major, and grade point average (GPA). The chosen strata are usually characteristics or attributes thought to be related to the dependent variables under study.

The particular population from which a sample is *actually* drawn is called a **sampling frame**, and it must be specified clearly. Technically speaking, all elements that are part of the population of interest to the research question should be part of the sampling frame. If they are not, any data collected may not be representative of the population. Often, however, sampling frames are incomplete, as the following example illustrates.

Suppose a researcher evaluates community opinion about snow removal by interviewing every fifth adult entering a local supermarket. The sampling frame would consist of all adults entering the supermarket while the researcher was standing

[5] This discussion of terms used in sampling is drawn primarily from Earl R. Babbie, *Survey Research Methods* (Belmont, Calif.: Wadsworth, 1973), 79–81.

outside. This sampling frame could hardly be construed as including all adult members of the community unless all adult members of the community made a trip to the supermarket *when* the researcher was there. Furthermore, use of such a sampling frame would probably introduce bias into the results. Perhaps many of the people who stayed at home rather than going to the supermarket considered the trip too hazardous because of poor snow removal. The closer the sampling frame is to the population of interest or theoretical population, the better.

Sometimes lists of elements exist that constitute the sampling frame. For example, a university may have a list of all students, or the Conference of Mayors may have a list of current mayors of cities with 50,000 residents or more. The existence of a list may be enticing to a researcher, since it removes the need to create one from scratch. But lists may represent an inappropriate sampling frame if they are out of date, incorrect, or do not really correspond to the population of interest. A common example would be if a researcher used a telephone directory as the sampling frame for interviewing sample households within the service area. Households with unlisted numbers would be missed, some numbers would belong to commercial establishments or no longer be working, and recently assigned numbers would not be included. Consequently, the telephone book could constitute an inaccurate or inappropriate sampling frame for the population in that area. Researchers should carefully check their sampling frames for potential omissions or erroneously included elements. Consumers of research should also carefully examine sampling frames to see that they match the populations researchers claim to be studying.

An example of a poll that relied on an incomplete sampling frame is the infamous *Literary Digest* poll of 1936. Despite being based on a huge sample, it predicted that the winner of the presidential election would be Alf Landon, not Franklin D. Roosevelt. This poll relied on a sample drawn from telephone directories and automobile registration lists compiled by the investigators. At that time telephone and automobile ownership were not as widespread as they are today. Thus, the sampling frame overrepresented wealthy individuals.[6] The problem was compounded by the fact that in the midst of the Great Depression an unprecedented number of poor people voted, and they voted overwhelmingly for Roosevelt, the eventual landslide winner.

A newer problem with the use of telephone directories is that in addition to some households not having phones, so many people have unlisted numbers or cell phones only that reliance on a printed list will quite possibly lead to a biased sample. In many instances a list of the complete population may not exist, or it may not be feasible to create one. It may be possible, however, to make a list of groups. Then the researcher could sample this list of groups and enumerate the elements only in those groups that are selected. In this case, the initial sampling frame would consist of a list of groups, not elements. For example, suppose you wanted to collect data on the attitudes and behavior of civic and social service volunteers in a large

[6] Ibid., 74–75.

metropolitan area. Rather than initially developing a list of all such volunteers—a laborious and time-consuming task—you could develop a list of all organizations known to use volunteers. Next, a subset of these organizations could be selected, and then a list of volunteers could be obtained for only this subset. (This process is called cluster sampling, which is discussed in greater detail below.)

A **sampling unit** is an entity listed in a sampling frame. In simple cases the sampling unit is the same as an element. In more complicated sampling designs it may be a collection of elements. In the previous example, organizations are the sampling units.

Types of Samples

Researchers make a basic distinction among types of samples according to how the data are collected. We mentioned earlier that political scientists often select a sample, collect information about elements in the sample, and then use those data to make inferences about the population from which the sample was drawn. In other words, they make inferences about the whole population from what they know about a smaller group. If a sampling frame is incomplete or inappropriate, **sample bias** will occur. In such cases the sample will be unrepresentative of the population of interest, and inaccurate conclusions about the population may be drawn. Sample bias may also be caused by a biased selection of elements, even if the sampling frame is a complete and accurate list of the elements in the population.

Suppose that in the survey of opinion on snow removal mentioned earlier every adult in the community did enter the supermarket while the researcher was there. And suppose that instead of selecting every fifth adult who entered, the researcher avoided individuals who appeared to be in a hurry or in poor humor (perhaps because of snowy roads). In this case the researcher's sampling frame was fine, but the sample itself would probably be biased and not representative of public opinion in that community. Because of the concern over sample bias, it is important to distinguish between two basic types of samples: probability and nonprobability samples. A **probability sample** is simply a sample for which each element in the total population has a known probability of being included in the sample. This knowledge allows a researcher to calculate how accurately the sample reflects the population from which it is drawn. By contrast, a **nonprobability sample** is one in which each element in the population has an unknown probability of being selected. Not knowing the probabilities of inclusion rules out the use of statistical theory to make inferences. Whenever possible, probability samples are preferred to nonprobability samples.

In the next several sections we consider different types of probability samples: simple random samples, systematic samples, stratified samples (both proportionate and disproportionate), cluster samples, and telephone samples. We then examine nonprobability samples and their uses.

Simple Random Samples

In a **simple random sample** each element and combination of elements has an equal chance of being selected. A list of all the elements in the population must be available, and a method of selecting those elements must be used that ensures that each element has an equal chance of being selected.[7] We review two common ways of selecting a simple random sample so that you can see how elements are given an equal chance of selection.

Note first that despite the seeming simplicity, it can be quite difficult in practice to draw a truly simple random sample. Try writing down one hundred (much less one thousand) random integers. If you are like most people, the chances are that subtle patterns will creep into the list. You may subconsciously, for example, have a slight predilection for sevens, in which case your list will contain too many of them and too few of other numbers. This is not just an academic issue but a practical problem that confronts researchers in all fields.

We explain a few alternative methods of drawing random samples. One way of selecting elements at random from a list is by assigning a number to each element in the sample frame and then using a random numbers table, which is simply a list of random numbers, to select a sample of numbers. A computer can also create random numbers for this purpose. However it is done, those units having the chosen numbers associated with them are included in the sample.

Suppose, for instance, we have a population of 3,000 elements and wish to draw from it a sample of 150. First number each member of the population, 1, 2, 3, and so on, up to 3,000. Then we can start at a random place in a random numbers table and look across and down the columns of numbers to identify our selections. Today, computers are typically used to create random numbers (see table 7-1). Each time a number between 0001 and 3000 appears, the element in the population with that number is selected. If a number appears more than once, that number is ignored after the first time, and we simply go on to another number. (This is called sampling without replacement.) For example, if we combine the adjacent cells of the first two columns in table 7-1 (a table of random integers), we would have the following, random numbers: 4633, 2339, 9816, 2038, and 0869. Because 0869, 2038, and 2339 fall between 0001 and 3000, they (or more precisely, the elements to which they are assigned) would be included in the sample. Doing the same for the next two columns,

[7] When used to describe a type of sample, *random* does not mean haphazard or casual; rather, it means that every element has a known probability of being selected. Strictly speaking, to ensure an equal chance of selection, *replacement* is required—putting each selected element back on the list before the next element is selected. In *simple* random sampling, however, elements are selected without replacement. This means that on each successive draw, the probability of an element's being selected increases because fewer and fewer elements remain. But for each draw, the probability of being selected is equal among the remaining elements. If the sample size is less than one-fifth the size of the population, the slight deviation from strict random sampling caused by sampling without replacement is acceptable. See Hubert M. Blalock Jr., *Social Statistics*, 2nd ed. (New York: McGraw-Hill, 1972), 513–14.

TABLE 7-1 **Fifty Random Numbers**

46	33	35	65	86	18	16	15	43	77
23	39	49	87	40	97	45	85	63	23
98	16	97	48	06	86	93	11	07	24
20	38	05	54	41	28	32	55	29	93
08	69	12	40	80	32	45	85	33	35

Note: The fifty pseudo-random integers lie between 0 and 99 and were computer-generated.

we would include elements 0554 and 1240. As long as we do not deliberately look for a certain number, we may start anywhere in the table and use any system to move through it. As we suggested earlier, it would not be acceptable to generate four-digit numbers in one's head, however, since the numbers would likely be biased in some way. Of course, for a real project we would automate the entire process by having a computer select the 150 random numbers that meet our criterion.

As another example, suppose we wanted to analyze the voting behavior of Supreme Court justices using a database that contains information on 172 men and women who have been nominated to serve on the Supreme Court since 1789. (Several individuals were nominated more than once, but for now we ignore this problem.) We treat this pool of subjects as a population. Suppose we want a sample of ten nominees.

A computer pseudo-random number generator spit out these numbers: 12 165 121 60 54 74 132 76 46 159. Hence, we would select the 46th, 165th, 121st, . . . nominee from the list for analysis. Thus, Samuel Chase and David Souter would be the first two nominees picked for the study.

Simple random sampling requires a list of the members of the population. Whenever an accurate and complete list of the target population is available and is of manageable size, a simple random sample can usually be drawn. For example, a random sample of members of Congress could be drawn from a list of all 100 senators and 435 representatives. A simple random sample of countries could be chosen from a list of all the countries in the world, or a random sample of American cities with more than 50,000 people could be selected from a list of all such cities in the United States. The problem, as we will see, is that obtaining such a list is not always easy or even possible.

Systematic Samples

Assigning numbers to all elements in a list and then using random numbers to select elements may be a cumbersome procedure. Fortunately, a **systematic sample**, in which elements are selected from a list at predetermined intervals, provides an alternative method that is sometimes easier to apply. It too requires a list of the target population. But the elements are chosen from the list systematically rather than randomly. That is, every *k*th element on the list is selected, where *k* is the number that will result in the desired number of elements being selected. This number is called the **sampling interval**, or the "skip" or number of elements between elements that are drawn and is simply $k = N/n$, where N is the "population" size and n is the desired sample size.

Go back to the Supreme Court nominees. We could treat the database as a list with 172 entries. If we wanted a sample of size $n = 10$, we would divide the total by 10 to get the sampling fraction or interval: $k = 172/10 \approx 17$. So starting at a random point we could take every 17th name. (If we started at 11, we would include the 11th, 28th, 45th, . . . nominees.)

Systematic sampling is useful when dealing with a long list of population elements. It is often used in product testing. Suppose you have been given the job of ensuring that a firm's tuna fish cans are sealed properly before they are delivered to grocery stores. And assume that your resources permit you to test only a sample of tuna fish cans rather than the entire population of tuna fish cans. It would be much easier to systematically select every 300th tuna fish ean as it rolls off the assembly line than to collect all the cans in one place and randomly select some of them for testing.

Despite its advantages, systematic sampling may result in a biased sample in at least two situations.[8] One occurs if elements on the list have been ranked according to a characteristic. In that situation the position of the random start will affect the average value of the characteristic for the sample. For example, if students were ranked from the lowest to the highest GPA, a systematic sample with students 1, 51, and 101 would have a lower GPA than a sample with students 50, 100, and 150. Each sample would yield a GPA that presented a biased picture of the student population.

The second situation leading to bias occurs if the list contains a pattern that corresponds to the sampling interval. Suppose

TABLE 7-2 An Abbreviated List of Supreme Court Nominees, 1787–2011

No.	Nominee	Birth Year
1	Jay, John	1745
2	Rutledge, John	1739
3	Cushing, William	1732
4	Harrison, Robert H.	1745
5	Wilson, James	1742
6	Blair, John, Jr.	1732
7	Iredell, James	1751
8	Johnson, Thomas	1732
9	Paterson, William	1745
10	Rutledge, John	1739
11	Cushing, William	1732
12	**Chase, Samuel**	**1741**
. . .	. . .	. . .
162	Scalia, Antonin	1936
163	Bork, Robert H.	1927
164	Kennedy, Anthony McLeod	1936
165	**Souter, David H.**	**1939**
166	Thomas, Clarence	1948
167	Ginsburg, Ruth Bader	1933
168	Breyer, Stephen G.	1938
169	Roberts, John G., Jr.	1955
170	Miers, Harriet E.	1945
171	Alito, Samuel A., Jr.	1950
172	Sotomayor, Sonia Maria	1954

Source: Lee Epstein, Thomas G. Walker, Nancy Staudt, Scott A. Hendrickson, and Jason M. Roberts, U.S. Supreme Court Justices Database. Accessed January 26, 2010. Available at http://epstein.law.north western.edu/research/justicesdata.html

Note: Duplicate names deleted.

[8] Hubert M. Blalock, Jr., *Social Statistics*, 2nd ed. (New York: McGraw-Hill, 1972), 515.

you were conducting a study of the attitudes of children from large families and you were working with a list of the children listed by age in each family. If the families included in the list all had six children and your sampling interval was six (or any multiple of six), then systematic sampling would result in a sample of children who were all in the same position among their siblings. If attitudes varied with birth order, then your findings would be biased.

Stratified Samples

A **stratified sample** is a probability sample in which elements sharing one or more characteristics are grouped and elements are selected from each group in proportion to the group's representation in the total population. Stratified samples take advantage of the principle that the more homogeneous the population, the easier it is to select a representative sample from it. Also, if a population is relatively homogeneous, the size of the sample needed to produce a given degree of accuracy will be smaller than for a heterogeneous population. In stratified sampling, sampling units are divided into strata with each unit appearing in only one stratum. Then a simple random sample or systematic sample is taken from each stratum.

A stratified sample can be either proportionate or disproportionate. In proportionate sampling, a researcher uses a stratified sample in which each stratum is represented in proportion to its size in the population—what researchers call a **proportionate sample**. For example, let's assume we have a total population of 500 colored balls: 50 each of red, yellow, orange, and green and 100 each of blue, black, and white. We wish to draw a sample of 100 balls. To ensure a sample with each color represented in proportion to its presence in the population, we first stratify the balls according to color. To determine the number of balls to sample from each stratum, we calculate the **sampling fraction,** which is the size of the desired sample divided by the size of the population. In this example, the sampling fraction is 100/500, or one-fifth of the balls. Therefore, we must sample one-fifth of all the balls in each stratum.

Since there are 50 red balls, we want one-fifth of 50 or 10 red balls. We could select these 10 red balls at random or select every fifth ball with a random start between 1 and 5. If we followed this procedure for each color, we would end up with a sample of 10 each of red, yellow, orange, and green balls and 20 each of blue, black, and white balls. Note that if we selected a simple random sample of 100 balls, there is a finite chance (albeit slight) that all 100 balls would be blue or black or white. Stratified sampling guarantees that this cannot happen, which is why stratified sampling results in a more representative sample. Some deviation from proportional representation will occur, however, depending on the sampling interval, the random start, and the number of sampling units in a stratum.

In selecting characteristics on which to stratify a list, you should choose characteristics that are expected to be related to or affect the dependent variables in your study. If you are attempting to measure the average income of households in a city, for example, you might stratify the list of households by education, sex, or race of household head. Because income may vary by education, sex, or race, you would want to make sure that the sample is representative with respect to these factors. Otherwise the sample estimate of average household income might be biased.

If you were selecting a sample of members of Congress to interview, you might want to divide the list of members into strata consisting of the two major parties, or the length of congressional service, or both. This would ensure that your sample accurately reflected the distribution of party and seniority in Congress. Some lists may be inherently stratified. Telephone directories are stratified to a degree by ethnic groups, because certain last names are associated with particular ethnic groups. Lists of Social Security numbers arranged consecutively are stratified by geographical area, because numbers are assigned based on the applicant's place of residence.

In the examples of stratified sampling we have considered so far, we assured ourselves of a more representative sample in which each stratum was represented in proportion to its size in the population. There may be occasions, however, when we wish to take a **disproportionate sample**. In such cases, we would use a stratified sample in which elements sharing a characteristic are underrepresented or overrepresented in our sample.[9]

For example, suppose we are conducting a survey of 200 students at a college in which there are 500 liberal arts majors, 100 engineering majors, and 200 business majors for a total of 800 students. If we sampled from each major (the strata) in proportion to its size, we would have 125 liberal arts majors; 25 engineering majors, and 50 business majors. If we wished to analyze the student population as a whole, this would be an acceptable sample. But if we wished to investigate some questions by looking at students in each major separately, we would find that 25 engineering students were too small a sample from which to draw inferences about the population of engineering students.

To get around this problem, we could sample disproportionately—for example, we could include 100 liberal arts majors, 50 engineering majors, and 50 business majors in our study. Then we would have enough engineering students to draw inferences about the population of engineering majors. The problem now becomes evaluating the student population as a whole, since our sample is biased due to an undersampling of liberal arts majors and an oversampling of engineering majors. Suppose engineering students have high GPAs. Our sample estimate of the student

[9] There are two reasons to use disproportionate sampling in addition to obtaining enough cases for statistical analysis of subgroups: the high cost of sampling some strata and differences in the heterogeneity of some strata that result in differences in sampling error. A researcher might want to minimize sampling when it is costly or increase sampling from heterogeneous strata while decreasing it from homogeneous strata. See Hubert M. Blalock, *Social Statistics*, 2nd ed. (New York: McGraw-Hill, 1972), 513–14, 518–19.

body's GPA would be biased upward because we have oversampled engineering students. Therefore, when we wish to analyze the total sample, not just students in a particular major, we need some method of adjusting our sample so that each major is represented in proportion to its real representation in the total student population.[10]

Table 7-3 shows the proportion of the population of each major and the mean GPA for each group in a hypothetical sample of college students. To calculate an unbiased estimate of the overall mean GPA for the college, we could use a **weighting factor,** a mathematical factor used to make a disproportionate sample representative. In this example, we would multiply the mean GPA for each major by the proportion of the population of each major (that is, the weighting factor).[11] Thus, the mean GPA would be .625(2.5) +.125(3.3) + .25(2.7) = 2.65.

Disproportionate stratified samples allow a researcher to represent more accurately the elements in each stratum and ensure that the overall sample is an accurate representation of important strata within the target population. This is done by weighting the data from each stratum when the sample is used to estimate characteristics of the target population. Of course, to accomplish disproportionate stratified sampling, the proportion of each stratum in the target population must be known.

Cluster Samples

Thus far, we have considered examples in which a list of elements in the sampling frame exists. There are, however, situations in which a sample is needed but no list of elements exists and to create one would be prohibitively expensive. A **cluster sample** is a probability sample in which the sampling frame initially consists of clusters of elements. Since only some of the elements are to be selected in a sample, it is unnecessary to be able to list all elements at the outset.

In cluster sampling, groups or clusters of elements are identified and listed as sampling units. Next, a sample is drawn from this list of sampling units. Then, for the sampled units only, elements are identified and sampled. For example, suppose we wanted to take an opinion poll of 1,000 persons in a city for which there is no complete list of city residents. We might begin by obtaining a map of the city and identifying and listing all blocks. This list of blocks becomes the sampling frame from which a small number of blocks are sampled at random or systematically. (The individual blocks are sometimes called the *primary sampling units.*) Next, we would go to the selected blocks and list all the dwelling units in those blocks. Then a sample of dwelling units would be drawn from each block. Finally, the households in the sampled dwellings would be contacted, and someone in each household would

[10] Ibid., 521–22.

[11] We could have obtained the same results by multiplying the GPA of each student by the weighting factor associated with the student's major and then calculating the mean GPA for the whole sample.

be interviewed for the opinion poll. Suppose there are 500 blocks and, from these 500 blocks, 25 are chosen at random. On these 25 blocks, a total of 4,000 dwelling units or households are identified. One-quarter of these households will be contacted because a sample of 1,000 individuals is desired. These 1,000 households could be selected with a random sample or a systematic sample.

Note that even though we did not know the number of households ahead of time, each household has an equal chance of being selected. The probability that any given household will be selected is equal to the probability of one's block being selected times the probability of one's household being selected, or 25/500 × 1,000/4,000 = 1/80. Thus, cluster sampling conforms to the requirements of a probability sample.

Our example involved only two samples or levels (the city block and the household). Some cluster samples involve many levels or stages and thus many samples. For example, in a national opinion poll, the researcher might list and sample states, list and sample counties within states, list and sample municipalities within counties, list and sample census tracts within municipalities, list and sample blocks within census tracts, and finally list and sample households—a total of six stages.

An advantage of cluster sampling is that it allows researchers to get around the problem of acquiring a list of elements in the target population. Cluster sampling also reduces fieldwork costs for public opinion surveys, because it produces respondents who are close together. For example, in a national opinion poll, respondents will not come from every state. This reduces travel and administrative costs.

Systematic, stratified (both proportionate and disproportionate), and cluster samples are acceptable and often more practical alternatives to the simple random sample. In each case, the probability of a particular element's being selected is known; consequently, the accuracy of the sample can be determined. The type of sample chosen depends on the resources a researcher has available and the availability of an accurate and comprehensive list of the elements in a well-defined target population.

TABLE 7-3 **Stratified Sample of Student Majors**

	Liberal Arts	Engineering	Business	Total
Number of students	500	100	200	800
Proportion or weight	.625	.125	.25	1.00
Size of sample	100	50	50	200
Sample mean grade point average	2.5	3.3	2.7	2.65

Note: Hypothetical data.

HELPFUL HINTS

Sampling in the Real World

Later in the book, we analyze data from *United States Citizenship, Involvement, Democracy (CID) Survey, 2006,* a nationwide study of political participation. The projected employed a multistage sampling design. A reconstruction of the steps taken to identify and interview respondents shows just how complex (and arduous) sampling can be:[12]

1. *Population:* "Eligible respondents were household members, males or females, age 18 years old and older. . . . The sample was designed to specifically represent the adult population residing in occupied residential housing units, and by definition excluded residents of institutions, group quarters, or those residing on military bases."
2. *Sampling frame:* All residential units.
3. *Stratification levels:* The four standard census regions and metropolitan areas.
4. *Clusters:* "Within each primary stratum, all counties, and by extension every census tract, block group and household, were ordered in a strict hierarchical fashion. . . . Within each metropolitan stratum, MSAs [metropolitan statistical areas] and their constituent counties were arrayed by size (i.e., number of households). Within each MSA, the central-city county or counties were listed first, followed by all non-central-city counties. In the four non-metropolitan strata, states and individual counties within each state were arrayed in serpentine order, North-to-South, and East-to-West. Within county, Census Tracts and Block Groups were arrayed in numerical sequence, which naturally groups together households within cities, towns, and other minor civil divisions (MCDs)."
5. *Selecting households:* "Within each sample PSU [primary sampling unit], two block groups (BG) were selected at random, without replacement. . . . All residential housing units within a sample BG were then

[12] Marc M. Howard, James L. Gibson, and Dietlind Stolle, *United States Citizenship, Involvement, Democracy (CID) Survey, 2006* (Ann Arbor, Mich.: Inter-University Consortium for Political and Social Research, 2007).

(Continued)

(Continued)

identified using the U.S. Postal Service Delivery Sequence File (DSF) and one address selected at random. The next fourteen residential addresses were then identified, along with any intervening commercial, vacant, or seasonal units. The result was a designated walking list that was supplied to each interviewer, along with a map showing the exact segment location, streets, addresses, etc. The street/address listing typically captures about 98% of all occupied housing units."

6. *Interviewer instructions:* "Interviewers were given street/address listings with 15 addresses, and were instructed to work the first ten pieces to a maximum of six callbacks. In order to properly manage the release of sample and strive to work all released sample to its maximum attempts, interviewers were asked to check in once they had attained five interviews or worked the first ten pieces to final dispositions or six active attempts, whichever came first. Throughout the field period the field director made daily decisions regarding whether each interviewer should continue working their first ten pieces or be provided more sample to work. Again, the overall goal was to attain a maximum number of attempts with as little sample as possible within a limited field period and an overall goal of approximately 1,000 completed interviews."

It is perhaps clear now why phone and Internet surveys are so popular.

Check out more Helpful Hints at **edge.sagepub.com/johnson8e**

Nonprobability Samples

A nonprobability sample is a sample for which each element in the total population has an unknown probability of being selected. Probability samples are usually preferable to nonprobability samples because they represent a large population fairly accurately and it is possible to calculate how close an estimated characteristic is to the population value. In some situations, however, probability sampling may be too expensive to justify (in exploratory research, for example), or the target population may be too ill-defined to permit probability sampling (this was the case with the television commercials example discussed earlier). Researchers also may feel that they can learn more by studying carefully selected and perhaps unusual cases than by studying representative ones. A brief description follows of some of the types of nonprobability samples.

With a judgmental sample, a researcher exercises considerable discretion over what observations to study, because the goal is typically to study a diverse and usually limited number of observations rather than to analyze a sample representative of a larger target population. Richard F. Fenno Jr.'s *Home Style,* which describes the behavior of eighteen incumbent representatives, is an example of research based on a purposive sample.[13] Likewise, a study of journalists that concentrated on prominent journalists in Washington, D.C., or New York City would be a purposive rather than a representative sample of all journalists.

A **quota sample** is a sample in which elements are sampled in proportion to their representation in the population. In this respect, quota sampling is similar to proportionate stratified sampling. The difference between quota sampling and stratified sampling is that the elements in the quota sample are not chosen in a probabilistic manner. Instead, they are chosen in a purposive or convenient fashion until the appropriate number of each type of element (quota) has been found. Because of the lack of probability sampling of elements, quota samples are usually biased estimates of the target population. Even more important, it is impossible to calculate the accuracy of a quota sample.

A researcher who decided to conduct a public opinion survey of 550 women and 450 men and who instructed his interviewers to select whomever they pleased until these quotas were reached would be drawing a quota sample. A famous example of an error-ridden quota sample is the 1948 Gallup Poll that predicted that Thomas Dewey would defeat Harry Truman for president.[14]

In a "snowball sample," respondents are used to identify other persons who might qualify for inclusion in the sample.[15] These people are then interviewed and asked to supply appropriate names for further interviewing. This process is continued until enough persons are interviewed to satisfy the researcher's needs. Snowball sampling is particularly useful in studying a relatively select, rare, or difficult-to-locate population such as draft evaders, political protesters, or even home gardeners who use sewage sludge on their gardens—a group estimated to constitute only 3 to 4 percent of households.[16]

[13] Richard F. Fenno Jr., *Home Style: House Members in Their Districts* (Boston: Little, Brown, 1978).

[14] Earl R. Babbie, *Survey Research Methods* (Belmont, Calif.: Wadsworth, 1973), 75.

[15] Snowball sampling is generally considered to be a nonprobability sampling technique, although strategies have been developed to achieve a probability sample with this method. See Kenneth D. Bailey, *Methods of Social Research* (New York: Free Press, 1978), 83. The "reputational approach" discussed in chapter 2 could be considered an example of this type of sample.

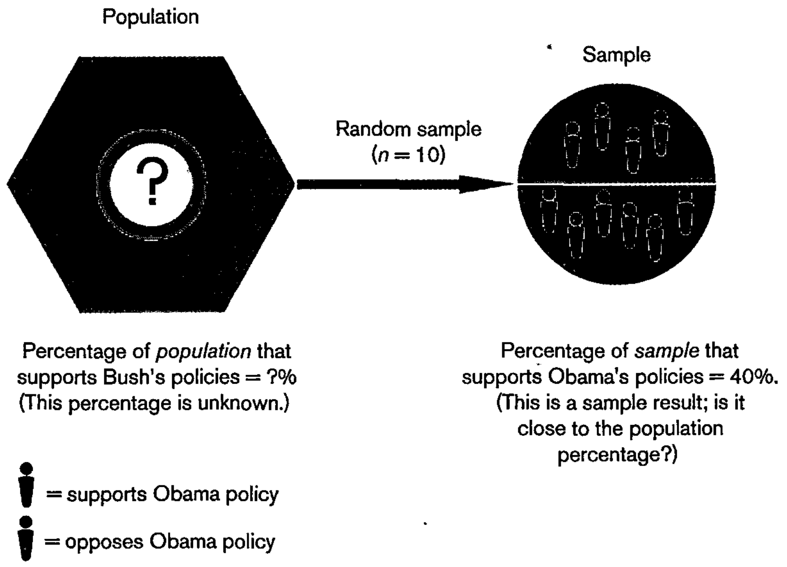
[16] Jane W. Bergsten and Stephanie A. Pierson, "Telephone Screening for Rare Characteristics Using Multiplicity Counting Rules," in *1982 Proceedings of the Section on Survey Research Methods* (Alexandria, Va.: American Statistical Association, 1982), 145–50. Available at http://www.amstat.org/sections/srms/proceedings/

We have discussed the various types of samples that political science researchers use in their data collection. Samples allow researchers to save time, money, and other costs. However, this benefit is a mixed blessing for by avoiding these costs, researchers must rely on information that is less accurate than if they had collected data on the entire target population. Now we consider the type of information that a sample provides and the implications of using this information to make inferences about a target population.

What Can Be Learned from a Sample of a Population

Suppose we want to measure support for President Obama's handling of the "war on terror." Figure 7-2 illustrates our problem. On one hand, at any given time a presumably *unknown* proportion of Americans back the president's policies, but we have little or no idea what that percentage is. (In an earlier section, we called this percentage a population *parameter.*) Imagine we were to draw (at random) a sample of ten adult Americans and count the number who are supportive. (Look at the right side of Figure 7-2.) Here we see that four out of ten, or 40 percent, of the respondents are supportive. This number, the sample statistic or estimator, provides an

FIGURE 7-2 **The Problem of Inference**


Population
Sample
Random sample
($n = 10$)
?
Percentage of *population* that supports Bush's policies = ?% (This percentage is unknown.)
Percentage of *sample* that supports Obama's policies = 40%. (This is a sample result; is it close to the population percentage?)
= supports Obama policy
= opposes Obama policy

estimate of the population proportion. Not having any other information, we might take it as our best approximation of public opinion on the matter. But just how good is it? Can we really say anything about the attitudes of millions of Americans based on a sample of just ten people? Before making a judgment, let's examine sampling and inference in a bit more detail.

Samples provide only estimates or approximations of population attributes. Occasionally these estimates may be right on the money. Most of the time, however, they will differ from the true value of the population parameter. When we report a sample statistic, we always assume there will be a margin of error, or a difference between the reported and actual values. For example, a finding that 53 percent of a random sample backs the president's goals and policies does not mean that exactly 53 percent of the public is sympathetic. It means merely that *approximately* 53 percent are. In other words, researchers sacrifice some precision whenever they rely on samples instead of enumerating and measuring the entire population. How much precision is lost (that is, how accurate the estimate is) depends on how the sample has been drawn and its size.

Where does the loss of precision or accuracy come from? The answer is chance, or luck of the draw. If you flip a coin ten times, you probably won't get exactly five heads, even if the coin is fair or the probability of heads is one-half. Randomness seems to be an innate feature of nature, at least on the scale at which we observe it. Just as with our coin toss, a random sample of ten (or even much larger) is not likely to produce precisely the value of a corresponding population parameter. But if we follow proper procedures and certain assumptions have been met (for example, the sample is a simple random sample from an infinite population), a sample statistic approximates the numerical value of a population parameter. If a population percentage really is 53, it is unlikely (*but not impossible*) for a sample result to be, say, 5 or 10 or 99 percent or some other "extreme" value. More likely, the sample estimate will be something like 40 or 60 percent. The difficulty is figuring out how far off the estimate is likely to be in any individual case. Here is where statistics helps.

The major goal of **statistical inference** is to make supportable conjectures about the unknown characteristics of a population based on sample statistics. The study of statistics partly involves defining much more precisely what *supportable* means. To make this clear, we introduce three concepts:

- Expected values
- Standard errors
- Sampling distributions

Although these terms may appear at first sight to have technical meanings, they can be given common-sense interpretations.

Expected Values

Let's look at a relatively simple example. A candidate for the state senate wants to know how many independents live in her district, which has grown rapidly in the past ten years. Although the Bureau of Elections reports that 25 percent of registered voters declined to name a party, she believes that the records are badly out of date. She asks you to conduct a poll to estimate the proportion of citizens, aged eighteen and older, who registered as independents rather than as Democrats or Republicans.

Suppose you interview ten *randomly* chosen adults living in the district and discover that two of them registered as independents.[17] Based on this finding, you could report that 20 percent of voters are registered as independents. Intuitively, however, you know that this estimate may be off by quite a bit, because you interviewed only ten people. The true proportion may be very different.

Now suppose for the moment that the Bureau of Elections' records are still accurate: one-fourth, or 25 percent, of the population is registered as independents, or, in more formal terms, $P = .25$, where P stands for the value of the population parameter. Of course, usually no one knows the population value because at the time of a poll it is unobserved, but we will pretend that we do in order to illustrate the ideas of sampling and inference. Your first estimate, .20, then, is a little bit below the true value. This difference is called the **sampling error**, which is the discrepancy between an observed and a true value that arises because only a portion of a population is observed.

What you need is some way to measure the amount of error or uncertainty in the estimate so that you can tell your client what the margin of error is. That is, you want to be able to say, "Yes, my estimate is probably not equal to the real value, but chances are that it is close." What exactly do words like *chances are* and *close* mean?

To answer those questions, imagine taking another, totally independent sample of ten adults from the same district and calculating the proportion of independents. (We will assume that not much time has passed since the first sample, so the probability of being an independent is still 25 percent.) This time the estimate turns out to be .30.

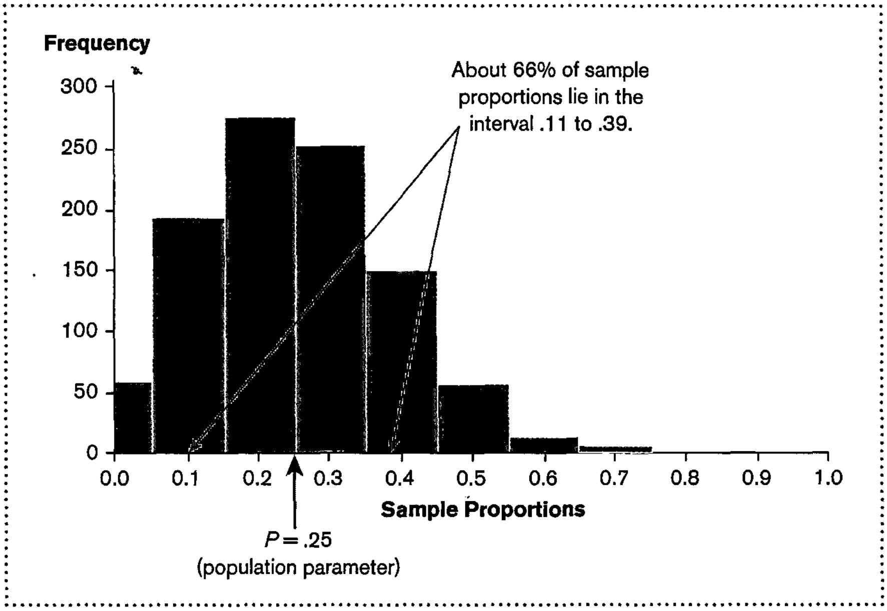
Repeating the procedure once more, you find that the next estimated proportion of independents is .40. This estimate, while quite high, is still possible. And after you take a fourth independent sample, you find that the estimated proportion, .15,

[17] The following remarks assume that we have a simple random sample, meaning (just as a reminder) that each member of the sample has been selected randomly and independently of all the others. We assume the same throughout the discussion in this section.

is again wide of the mark. So far, two of your estimates have been too large, two too low, and none exactly on target. But notice that the average of the estimates, (.20 +.30 +.40 +.15)/4 = .26, is not far from the real value of .25.

What would happen, you might wonder, if you repeated the process indefinitely? That is, what would happen if you took an infinite number of independent samples of $N = 10$ and calculated the proportion of independents in each one?[18] (Throughout this discussion, we use n to denote the size of a sample.) After a while, you would have an extended list of sample proportions or percentages. What would their distribution look like? Figure 7-3 gives an idea. In brief, we programmed a computer to take 1,000 samples (each of size 10) from a hypothetical population in which $P = .25$. This technique permits us to investigate the behavior of a huge number of sample outcomes.

FIGURE 7-3 **Distribution of 1,000 Sample Proportions (Sample Size = 10)**



Source: Simulated data.

Note: Mean of 1,000 sample proportions is .248.

[18] This procedure, called sampling with replacement, is premised on the assumption that, at least theoretically, people will sooner or later be interviewed twice or more. We ignore this nuance, because it does not affect the validity of the conclusions in this case.

The separate sample proportions are spread around the true value ($P = .25$) in a bell-curve-shaped distribution—that is, a curve with a single peak and more or less symmetric or equal tails. A few of the estimates are quite low, even close to zero, while a few more of them are way above .25. (The frequencies can be determined by looking at the *y*-axis, the vertical line.) Yet the vast majority is in the range .05 to .45, and the center of the distribution (the average of the 1,000 sample proportions) is near .25, the actual population value. Indeed, the average of the 1,000 proportions in this particular data set is .248, which lies very close to the true value! This is no coincidence, as we will see.

This illustration highlights an important point about samples and the statistics calculated from them. If statistics are calculated for each of many, many independently and randomly chosen samples, their average or mean will equal the corresponding true, or population, quantity, no matter what the sample size. Statisticians refer to this mean as the **expected value** (*E*) of the estimator. This idea can be stated more succinctly.

In the case of a sample proportion based on a simple random sample, we have

$$E(p) = P,$$

where *p* is the estimated proportion, and the equation reads, "The *expected* (or long-run, or average) value of sample proportions equals the population proportion, *P*."

In plain words, although any particular estimate result may not equal the parameter value of the population from which the data come,[19] if the sampling procedure were to be repeated an infinite number of times and a sample estimate calculated each time, then the average, or mean, of these results would equal the true value. This fact gives us confidence in the sampling method, though not in any particular sample statistic. Since Figure 7-3 includes only 1,000 estimates, not an infinite number, it only illustrates what can be demonstrated mathematically for many types of sample statistics.

Measuring the Variability of the Estimates: Standard Errors

Besides telling us the expected value for the population, statistical theory also tells us that sample proportions will fall above and below the true value in a predictable manner, as suggested by Figure 7-3. That is, there is variation or variability in the outcomes. As we just observed, most of the sample proportions fall between .05

[19] Indeed, in all likelihood it will not exactly equal the population value.

and .45 (or 5 and 45 percent). A few will be much larger or smaller, but they will be the exceptions. Consequently, we can use a graph like that shown in Figure 7-3 to determine approximately the likelihood of getting a particular sample result *if* the true value of the population from which the samples have been drawn is .25. For example, what are the chances of getting a sample proportion of .29 if the population proportion is .25? The answer: very likely. Why? Because statistics tells us that most sample results will be close to the true value. But suppose a sample proportion turns out to be .75. If the true number is .25, is this a likely result? Look at the figure. It suggests that a value that far from .25 occurs only rarely. So the answer might be, "A sample proportion of .75 is possible but not very probable." (You can use the areas in the rectangles to "guesstimate" the chances.)

The fact that statistics behave in this manner helps us make inferences. To anticipate the material in later chapters, let us continue to hypothesize that the true proportion is .25. Now assume that a sample of ten produces a proportion of .19. Given that such a result is reasonably possible—look at Figure 7-3 once again—we might conclude that this hypothesis cannot be rejected. However, if the sample result turned out to be .9, we would be justified in concluding that the hypothesis does not hold water and should be rejected. Why? Because it is very unlikely that we would get a sample result of .9 from a population where $P = .25$. Therefore, we're on pretty safe ground if we conclude that P is not .25 and has to be around .9.

Of course, we could be making a mistake. It is possible that the true proportion is .25 even though our sample estimate is way above that number. If we did reject the hypothesis (that is, $P = .25$), we would be wrong. Yet the chances of making this kind of error are relatively small. That's what people mean when they say they have confidence in an estimate. (Confidence does not equal certainty, just as in legal trials judgments are based on the standard of reasonable doubt, not absolute, infallible knowledge.)

The mathematical term for the variation around the expected value is the standard error of the estimator, or **standard error** for short. (You may know this term as "sampling error.") Loosely speaking, the standard error provides a numerical indication of the variation in our sample estimates. (Like all statistical indicators, it has its own symbol, $\hat{\sigma}$.) The standard error of your first poll of 10 adults is .13.[20] As of now, this number has no obvious meaning, but, as shown later, it can be used to make probability statements such as "roughly two-thirds of the sample proportions lie in the interval between .11 and .39." So now you can tell your client, the senator, that based on just the first (and presumably the only) sample you have taken, the true

[20] The standard error for proportions is calculated from the formula:

$$\hat{\sigma} = \sqrt{\frac{P(1-P)}{n}} = \sqrt{\frac{(.2)(.8)}{10}} = .13$$

proportion of independents in the district is probably somewhere between 7 and 33 percent, with the best bet being 20 percent. When she asks what you mean by *probably*, you can tell her, "I am about 66 percent sure." (You might be able to recognize this point by looking at the frequencies represented by the bars in the graph.)

Not surprisingly, your first estimate may not be very helpful to the campaign, which must decide how to target its limited resources. After all, if the percentage of independents in the district is as low as 7 percent, the senator might follow one strategy, but if it is 33 percent or more, she might do something else. As a result, the senator would like you to narrow the range of uncertainty. What can you do? The answer may be obvious: take a larger sample.

Imagine that you increase the sample size to 50 registered voters ($N = 50$). The population proportion (P) is still .25. Now note the estimated proportion. This time, you might find that 15 respondents out of 50, or 30 percent, are independents. Because of our omniscience, we know this estimate is a bit too large. But, as before, let us repeat the process. If you drew 1,000 independent random samples, each containing 50 observations, and plotted the distribution of the estimated proportions, you would get a graph similar to the one in Figure 7-4.

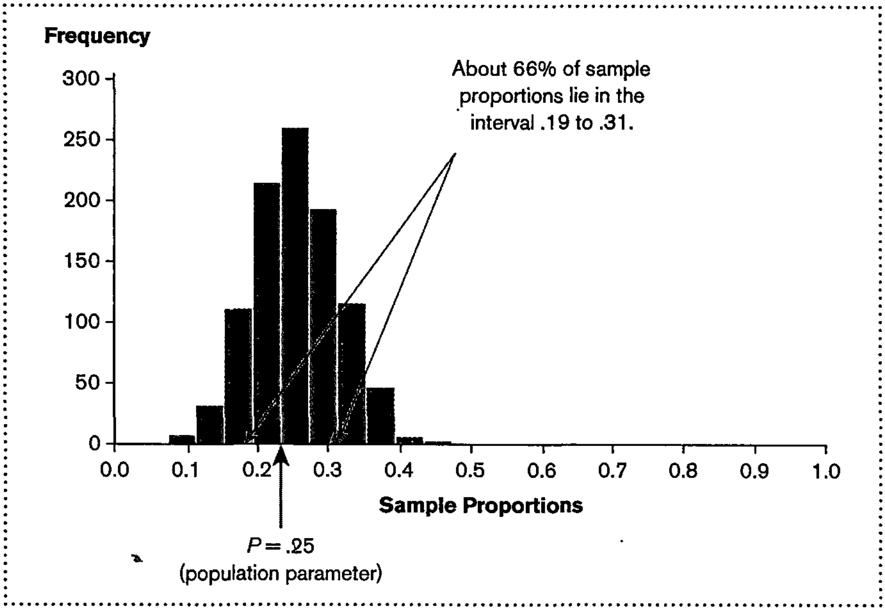
For the hypothetical data shown in Figure 7-4, the mean of the 1,000 sample proportions is .252, a value quite near the true number.[21] The figure illustrates once again what can be shown mathematically—namely, that the distribution of *p*s is approximately bell shaped, with the expected or long-run value of sample estimates being equal to the true proportion of the population from which the samples have been collected. Also, notice that the distribution is not as spread out as the one depicted in Figure 7-3. In our statistical language, the standard error is smaller, .06 now versus .14 previously. Hence, about two-thirds of the sample proportions fall in the interval .19 to .31, which is about half the width of the one based on ten cases; very few fall in the tails of the distribution. So increasing the sample size gives us more confidence that .252 is near the true value.

To cement the point, let us repeat the simulation using a much larger sample, $N = 500$. The result appears in Figure 7-5. It, too, shows that the average of the sample proportions is close to the true value and that the variability of the estimates, the sampling error, has been greatly reduced.

This finding illustrates the generalization that the sample size affects the magnitude of sampling variation: *the larger the sample, the smaller the standard error*. That statement, in turn, implies that as sample sizes increase, the range of sample estimators decreases. (This fact may be consistent with your intuition that large samples should be more "accurate" than small ones in the sense that as *n* increases so does the precision of the

[21] Note, too, that it is close to the value obtained from the 1,000 samples, where $N = 10$. So the average of the *p*s based on samples of 10 is not much different from the average based on samples of 50.

FIGURE 7-4 **Distribution of 1,000 Sample Proportions (Sample Size = 50)**



Source: Simulated data.

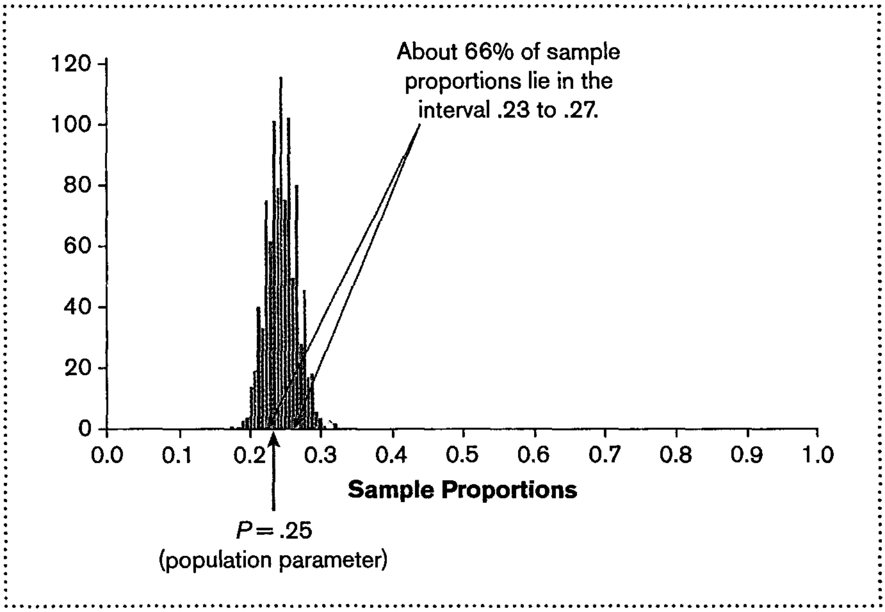
Note: Mean of 1,000 sample proportions is .252.

estimator.) But keep in mind that the expected value of sample estimators does not depend on the sample size. Instead, it is the confidence placed in them that does.

Sample Size

Table 7-4 summarizes our results for samples of size 10, 50, and 500. For instance, the row labeled "10" contains the results of taking 1,000 independent samples (each of $N = 10$) from a population in which $P = .25$. The average of the 1,000 sample proportions is .248, the standard error is .14, the interval containing about 66 percent of the sample proportions is .11 to .39, and the lowest and highest proportions are 0 and .70. Similarly, the next row contains the results of 1,000 samples of $N = 50$. For these sets of simulated data, the average of the sample proportions is always close to the true value no matter how large the sample, again illustrating the argument about expected values. But—and herein lies the crux of the argument—the measures of variability of the proportions decrease considerably as the sample sizes get larger. The numbers may seem small to you, but notice that the variability of the sample results based on 10 cases (.14) is more than twice as large as the corresponding number for the samples of size 50 (.06).

FIGURE 7-5 **Distribution of 1,000 Sample Proportions (Sample Size = 500)**



Source: Simulated data.

Note: Mean of 1,000 proportions is .250.

What does all this mean in plain English? Small sample sizes are not invalid or worthless. The expected values of many of their sample statistics will equal the population parameters. But confidence intervals based on small samples may be much too wide or imprecise to be useful.

We can illustrate the relationship between sample size and precision with still another example. Assume that we want to estimate a population mean, and suppose further that we want to be 99 percent certain about our estimate. (Notice that we have established a specific level of confidence—99 percent certainty.) To achieve this level of confidence, how wide off the mark can our estimate be and still be useful? Once we answer this question, we can choose an appropriate sample size. For example, if we want to say with 99 percent certainty that the interval $25,500 to $28,500 contains the true mean, then we would need a sample of a certain size (perhaps 200). But if we want to be 99 percent certain that the mean lies between $26,500 and $26,600—a mere $100 difference—then we will need a much larger sample.[22]

[22] Sample size is not the only factor that affects statistical inferences. For a somewhat advanced discussion, see Dennis D. Boos and Jacqueline M. Hughes-Oliver, "How Large Does *n* Have to Be for *Z* and *t* Intervals?" *American Statistician* 54, no. 2 (2000): 121–28.

TABLE 7-4 **Properties of Samples of Different Sizes**

Sample Size	66% Average (mean)[a]	Standard Error[b]	Confidence Interval	Minimum Proportion	Maximum Proportion	Range of Proportions
10	.248	.14	.11–.39	0	.70	.70
50	.252	.06	.19–.31	.1	.48	.38
500	.250	.02	.23–.27	.19	.32	.13

[a] Each mean is the average of 1,000 sample proportions taken from a population in which the true probability (the parameter of interest in this case) $P = .25$.

[b] This term measures the variation or variability of the sample proportions. It indicates the magnitude of sampling error.

Decisions about sample sizes involve trade-offs. Perhaps our state senate candidate wants an estimated proportion to be within 1 or 2 percent of the true value, but does she have sufficient funds to collect a large enough survey? If not, she might have to settle for a wider confidence interval.[23]

HELPFUL HINTS

How Large a Sample?

As we learned earlier, the key to controlling sampling error is the sample size. Generally, the larger the sample, the smaller the sampling error, as measured by the standard error. Given that sample size figures so prominently in **sampling distributions**, you might think that by increasing *N* you could reduce uncertainty to near zero. However, the relationship between sample size and sampling error is exponential rather than linear. For example, to cut sampling error in half, the sample size must be quadrupled. This means that researchers must balance the costs of increasing sample size with the size of the sampling error they are willing to tolerate.

Check out more Helpful Hints at **edge.sagepub.com/johnson8e**

[23] Sampling error also depends on the type of sample drawn. For a given sample size, a simple random sample provides a more accurate estimate of the target (that is, a smaller margin of error) than does a cluster sample. Sampling error is also smaller for an attribute that is shared by almost all elements in the sample than for one that is distributed across only half of the sample elements. Finally, sampling error is reduced if the sample represents a significant proportion of the target population—that is, if the sampling fraction is greater than one-fourth of the target population. Because this is unusual, however, the effect of the sampling fraction on sampling error is generally minuscule.

Conclusion

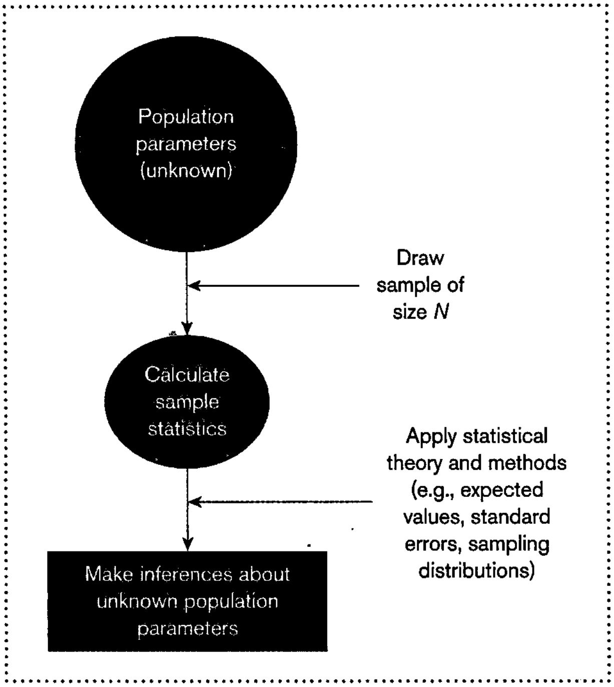
In this chapter, we discussed what it means to select a sample out of a target population, the various types of samples that political scientists use, and the kinds of information they yield. Figure 7-6 provides an intuitive summary of sampling in the research process.

The following guidelines may help researchers who are deciding whether or not to rely on a sample as well as students who are evaluating research based on sample data:

- If cost is not a major consideration, and the validity of the measures will not suffer, it is generally better to collect data for the complete target population than for just a sample of that population.
- If cost or validity considerations dictate that a sample be drawn, a probability sample is usually preferable to a nonprobability sample. The accuracy of sample estimates can be determined only for probability samples. If the desire to represent a target population accurately is not a major concern or is impossible to achieve, then a nonprobability sample may be used.
- Probability samples yield estimates of the target population. All samples are subject to sampling error. No sample, no matter how well drawn, can provide an exact measurement of an attribute of, or relationship within, the target population.

Fortunately, statistical theory gives us methods for making systematic inferences about unknown parameters and for objectively measuring the probabilities of making inferential errors. This information allows the researcher and the scientific community to judge the tenability of many empirical claims.

FIGURE 7-6 **The Process of Making Inferences from Samples**


Population parameters (unknown)
Draw sample of size *N*
Calculate sample statistics
Apply statistical theory and methods (e.g., expected values, standard errors, sampling distributions)
Make inferences about unknown population parameters

TERMS INTRODUCED

Cluster sample. A probability sample that is used when no list of elements exists. The sampling frame initially consists of clusters of elements.

Disproportionate sample. A stratified sample in which elements sharing a characteristic are underrepresented or overrepresented in the sample.

Element. A particular case or entity about which information is collected; the unit of analysis.

Estimator. A statistic based on sample observations that is used to estimate the numerical value of an unknown population parameter.

Expected value. The mean or average value of a sample statistic based on repeated samples from a population.

Nonprobability sample. A sample for which each element in the total population has an unknown probability of being selected.

Population. All the cases or observations covered by a hypothesis; all the units of analysis to which a hypothesis applies.

Population parameter. A characteristic or an attribute in a population (not a sample) that can be quantified.

Probability sample. A sample for which each element in the total population has a known probability of being selected.

Proportionate sample. A probability sample that draws elements from a stratified population at a rate proportional to the size of the samples.

Quota sample. A nonprobability sample in which elements are sampled in proportion to their representation in the population.

Sample. A subset of observations or cases drawn from a specified population.

Sample bias. The bias that occurs whenever some elements of a population are systematically excluded from a sample. It is usually due to an incomplete sampling frame or a nonprobability method of selecting elements.

Sample statistic. The estimator of a population characteristic or attribute that is calculated from sample data.

Sampling distribution. A theoretical (nonobserved) distribution of sample statistics calculated on samples of size *N* that, if known, permits the calculation of confidence intervals and the test of statistical hypotheses.

Sampling error. The difference between a sample estimate and a corresponding population parameter that arises because only a portion of a population is observed.

Sampling fraction. The proportion of the population included in a sample.

Sampling frame. The population from which a sample is drawn. Ideally, it is the same as the total population of interest to a study.

Sampling interval. The number of elements in a sampling frame divided by the desired sample size.

Sampling unit. The entity listed in a sampling frame. It may be the same as an element, or it may be a group or cluster of elements.

Simple random sample. A probability sample in which each element has an equal chance of being selected.

Standard error. The standard deviation or measure of variability or dispersion of a sampling distribution.

Statistical inference. The mathematical theory and techniques for making conjectures about the unknown characteristics (parameters) of populations based on samples.

Stratified sample. A probability sample in which elements sharing one or more characteristics are grouped and elements are selected from each group in proportion to the group's representation in the total population.

Stratum. A subgroup of a population that shares one or more characteristics.

Systematic sample. A probability sample in which elements are selected from a list at predetermined intervals.

Weighting factor. A mathematical factor used to make a disproportionate sample representative.

SUGGESTED READINGS

Govindarajulu, Zakkula. *Elements of Sampling Theory and Methods.* Upper Saddle River, N.J.: Prentice Hall, 1999.

Kish, Leslie. *Survey Sampling.* New York: John Wiley & Sons, 1995. (Originally published 1965.) This is the classic treatment of this subject.

Levy, Paul S., and Stanley Lemeshow. *Sampling of Populations: Methods and Applications.* 3rd ed. New York: Wiley, 1999.

Lohr, Sharon L. *Sampling: Design and Analysis.* Pacific Grove, Calif.: Duxbury Press, 1999.

Rea, Louis M., and Richard A. Parker. *Designing and Conducting Survey Research: A Comprehensive Guide.* 2nd ed. San Francisco: Jossey-Bass, 1997.

Rosnow, Ralph L., and Robert Rosenthal. *Beginning Behavioral Research: A Conceptual Primer.* 3rd ed. Upper Saddle River, N.J.: Prentice Hall, 1998.

CHAPTER 8

Making Empirical Observations

Firsthand Observation

CHAPTER OBJECTIVES

8.1 Identify different types of data and collection techniques.

8.2 Describe the role of direct observation as a data collection technique.

8.3 Explain the role of indirect observation in data collection.

8.4 Discuss the ethical issues concerning observation.

POLITICAL SCIENTISTS TEND TO USE three broad types of empirical observations, or data collection methods, depending on the phenomena they are interested in studying. In this chapter, and the two that follow, we will discuss them; we will discuss firsthand observation in this chapter, document analysis in chapter 9, and interview data in chapter 10. Each of these data collection methods brings its own unique advantages and disadvantages and many choices for researchers. We begin with an overview of the three approaches and begin a discussion of some of the issues that researchers must consider when choosing one data collection method over another.

Types of Data and Collection Techniques

Interview data, discussed in chapter 10, are derived from individuals. This type of data collection may involve interviewing a representative cross-section of the national adult population or a select group of political actors, such as leaders of nongovernmental organizations like the International Committee of the Red Cross. It may involve face-to-face interviews or interviews conducted over the phone or through the mail or Internet. Alternatively, researchers may use highly structured interviews in which a questionnaire is followed closely by less-structured, open-ended discussions. Regardless of the particular type of interview setting, however, the essentials of the data collection method are the same: the data come from responses to the verbal or written cues of the researcher, and the respondent knows these responses are being recorded.

In addition to interview data, political scientists use documents (newspapers, photographs, audio-visual clips, hearing testimony, press releases, letters, and diaries) as well as statistical data that exist in various archival records. We refer to these sources of data collectively as the written record. This type of data collection, known as document analysis (the subject of chapter 9), relies heavily on the record-keeping activities of government agencies, private institutions, interest groups, media organizations, and even private citizens. Also included in what we refer to as the written record are data collected first as interview data but then aggregated and reported in summary form for groups of individuals. For example, unemployment statistics are derived from the Census Bureau's Current Population Survey, a household survey conducted each month. What often sets document analysis apart from other data collection methods is that the researcher is usually not the original collector of the data and the original reason for the collection of the data may not have been to further a scientific research project.

Finally, this chapter discusses how data may be collected by making firsthand observations in a field study or in a laboratory setting. In firsthand observation the researcher collects data on political behavior by observing either the behavior itself (**direct observation**) or some physical trace of the behavior (**indirect observation**). It involves firsthand examination of activities, behavior, events, relationships, or the like. It may even involve observing and recording speech, but unlike interviewing, this method of data collection does not rely on people's verbal responses to verbal stimuli presented by the researcher.

Choosing among Data Collection Methods

A political scientist's choice of data collection method depends on many factors. One important consideration is the validity of the measurements that a particular method will permit. For example, a researcher who wants to measure the crime rate of different cities may feel that the crime rates reported by local police departments to the FBI are not sufficiently accurate to support a research project. The researcher may be concerned that some departments overreport and some underreport various criminal acts or that some victims of crimes may fail to report the crimes to the police, hence rendering that method of collecting data and measuring the crime rate unacceptable. Therefore, the researcher may decide that a more accurate indication of the crime rate can be attained by interviewing a sample of citizens in different cities and asking them how much crime they have experienced themselves.

Also reflecting a concern over the validity of measurements, Susan J. Carroll and Debra J. Liebowitz noted that scholars of women and politics have criticized the use of survey research to study the political participation of women.[1] One problem is that existing conceptions of what is considered "political," and hence what is asked about in survey questions, may not fully capture the range of women's political activity. Carroll and Liebowitz suggested that researchers look at the issue inductively—that is, study women's activities and determine in what ways their activities are political. For this approach, observation, in-depth interviews, and focus groups, rather than structured questionnaires, are more appropriate data collection methods.

A political scientist is also influenced by the **reactivity** of a data collection method—the effect of the data collection itself on the phenomena being measured. When people know their behavior is being observed and know or can guess the purpose of the observation, they may alter their behavior. As a result, the observed behavior may be an unnatural reaction to the process of being observed. People may be reluctant, for example, to admit to an interviewer that they hold views or engage in behavior that is unpopular or embarrassing or even immoral or illegal. Thus, many researchers prefer unobtrusive or nonreactive measures of political behavior, because they believe that the resulting data are less likely to include observations based on responses or behaviors that conceal true feelings, beliefs, or motives.

The population covered by a data collection method is another important consideration for a researcher. The population of interest determines whose behavior the researcher observes. One type of data may be available for only a few people, whereas another type may permit more numerous, interesting, and worthwhile comparisons. A researcher studying the behavior of political consultants, for example, may decide that relying on the published memoirs of a handful of consultants will not adequately cover the population of consultants (not to mention the validity problems of the data) and that it would be better to seek out a broad cross-section of consultants and interview them. Or a researcher interested in political corruption may decide that interviewing a broad cross-section of politicians charged with various corrupt practices is not feasible and that data (of a different kind) could

[1] Susan J. Carroll and Debra J. Liebowitz, "Introduction: New Challenges, New Questions, New Directions," in *Women and American Politics: New Questions, New Directions,* ed. Susan J. Carroll (Oxford, UK: Oxford University Press, 2003), 1–29.

be obtained for a more diverse set of corrupt acts from accounts published in the mass media.

Additionally, cost and availability are crucial elements in the choice of a data collection technique. Some types of data collection are simply more expensive than others, and some types of observations are made more readily than others. Large-scale interviewing, for example, is very expensive and time-consuming, and the types of questions that can be asked and behaviors that can be observed are limited. Although the costs of data generated through interviews or the written record may be high, the cost of firsthand observation through the expenditure of time (if the researcher does it) or money (if the researcher pays others to do it) will generally be even higher. Data from archival records are usually much less expensive, since the record-keeping entity has borne most of the cost of collecting and publishing the data. With the increased use of computers, many organizations are systematically collecting data of interest to researchers. A disadvantage, however, may be that the data must be made available by the record-keeping organization, which can refuse a researcher's request or take a long time to fill it.

Data collected through firsthand observation is an example of **primary data**—that is, data recorded and used by the researcher making the observations—whereas data from interviews or the written record can be primary data or **secondary data**—data used by a researcher who did not personally collect the data. Most data collected through direct observation are recorded in the form of personal notes, recordings, and transcripts. These data are less likely to be publicly available because notes, in particular, are highly individualized and intended to help the person taking the notes remember observations. Hence, they would be relatively dissatisfactory to others even if they were made publicly available.

The high cost of direct and indirect observation means that most students will not often have the resources to make their own observations for use in a research paper, except in the most limited fashion. Students will often find suitable data generated through interviews or the written record for free in publicly available data archives (see chapters 9 and 10), but students wishing to use data generated through direct or indirect observation must usually rely on their own ability to make the observations. For example, you might be able to use observations made during your internship with a political campaign in an analysis of election strategy, but you are not likely to have the time to make firsthand observations across multiple campaigns. In most cases, it will be more cost-effective to rely on other, more readily available sources of data for research projects.

Some students, however, working on a larger-scale project like a thesis, can use fieldwork to make insightful observations and conclusions. For example, Gina Yannitell Reinhardt spent a year in Brazil studying the Afrobrasilian political movement with the support of a grant for recent college graduates who had not yet enrolled in graduate school.[2] Others spend a semester abroad and take the opportunity to make observations of political life in another country.

In addition to these factors, researchers must consider the ethical implications of their proposed research. In most cases, the research topics you are likely to propose will not raise serious ethical concerns, nor will your choice of method of data collection hinge on the risk it may pose to human subjects. Nevertheless, you should be aware of the ethical issues and risks to others that can result from social science research, and you should be aware of the review process that researchers are required to follow when proposing research involving human subjects.

In this chapter and in chapters 9 and 10, the relative advantages and disadvantages of each of the major data collection methods are examined with respect to the factors of validity, reactivity, population coverage, cost, and availability. We also point out the ethical issues raised by some applications of these data collection methods.

Firsthand, Direct Observation

Social scientists have been making firsthand observations of human behavior since the beginning of the disparate social science disciplines. Firsthand observation includes both qualitative and quantitative methods, with deep roots in anthropology, psychology, and sociology in particular. Anthropologists have been making firsthand observations of human behavior for well over a century with a method called ethnography. **Ethnography** is generally used to go beyond description of events or actions to reveal the "cultural constructions, in which we live."[3] The goal is therefore to make cultural interpretation through personal observation of everyday life. The method is commonly characterized as one of "thick description" that captures as many details as possible[4] or, as Wedeen defined ethnography, "immersion in the place and lives of people under study."[5]

Ethnography has been adopted in other disciplines, including political science, and has taken on many different forms for different purposes. Political scientists have used firsthand observation to study democratization, political participation, social movements, political campaigning, community politics, program implementation, judicial proceedings, lawmaking, and other topics. In fact, any student who has had

[2] Gina Yannitell Reinhardt, "I Don't Know Monica Lewinsky, and I'm Not in the CIA. Now How about That Interview?" *PS: Political Science and Politics*, 42, no. 2 (2009): 295–98.

[3] Brian Hoey, "A Simple Introduction to the Practice of Ethnographic Fieldnotes," Marshall University Digital Scholar 1–10. Available at: http://works.bepress.com/brian_hoey/12

[4] C. Geertz, *The Interpretation of Cultures* (New York: Basic Books, 1973).

[5] Lisa Wedeen, "Reflections on Ethnographic Work in Political Science," *Annual Review of Political Science* 13, no. 1 (2010): 255–72.

an internship, kept a daily log or a diary, and written a paper based on his or her experiences has used this method of data collection.

Every day we "collect data" using observational techniques. We observe some attribute or characteristic of people and infer some behavioral trait from that observation. For example, we watch the car in front of us swerve between traffic lanes and conclude that the driver has been drinking. We observe the mannerisms, voice pitch, and facial expressions of a student making a presentation in one of our classes and decide that the person is exceptionally nervous. Or we decide that most of the citizens attending a public hearing are opposed to a proposed project by listening to their comments to each other before the start of the hearing. The observational techniques used by political scientists are only extensions of this method of data collection. They resemble everyday observations but are usually more self-conscious and systematic.

Firsthand observations may be classified in two basic categories: direct and indirect.[6] For example, a direct method of observing college students' favorite studying spots in classrooms and office buildings would involve walking around the buildings and recording students' locations. An indirect method of observing the same behavior would be to arrive on campus early in the morning before the custodial staff and measure the number of food wrappers, drink containers, and other pieces of debris at various locations. The vast majority of observation studies conducted by political scientists involve direct observation, in which the researcher observes actual behavior, with the observation more likely to occur in a field study that takes place in a natural setting than in a laboratory. The term **field study** is typically used to refer to open-ended and wide-ranging, rather than structured, observation in a natural setting like a home or office building, a community, a city, or even a country or region. In field studies, researchers typically ask questions of the people they are observing; thus, field studies also involve collection of interview data.

Direct Observation in a Natural Setting

Direct observation in natural settings has several advantages. One advantage of observing people in a natural setting is that people generally behave as they would ordinarily. Furthermore, the investigator is able to observe people for longer periods than would be possible in a laboratory. In fact, one of the striking features of field studies is the considerable amount of time an investigator may spend in the field. It is not uncommon for investigators to live in the community they are observing for a year or more. William F. Whyte's *Street Corner Society* was based on three years of observation (1937–1940), and Marc Ross's study of political participation in

[6] Eugene J. Webb, Donald T. Campbell, and Richard D. Schwarz, *Nonreactive Measures in the Social Sciences,* 2nd ed. (Boston: Houghton Mifflin, 1981).

Nairobi, Kenya, took more than a year of field observation.[7] To study the behavior of US representatives in their districts, Richard Fenno traveled intermittently for almost seven years, making thirty-six separate visits and spending 110 working days in eighteen congressional districts.[8] Ruth Horowitz spent three years researching youth in an inner-city Chicano community in Chicago.[9] Raphael Schlembach observed activists participating in the Camp for Climate Action in the United Kingdom over a period of four years.[10]

Sometimes researchers have no choice but to observe political phenomena as they occur in their natural setting. Written records of events may not exist, or the records may not cover the behavior of interest to the researcher. Relying on personal accounts of participants may be unsatisfactory because of participants' distorted views of events, incomplete memories, or failure to observe what is of interest to the researcher. Joan E. McLean suggested that researchers interested in studying the decision-making styles of women running for public office need to spend time with campaigns in order to gather information as decisions are being made, rather than rely on postelection questionnaires or debriefing sessions.[11]

You may look upon an internship, volunteer work, or participation in a community or political organization as an opportunity to conduct your own research using direct observation. More than likely, your research will be a case study in which you are able to compare the real world with theories and general expectations suggested in course readings and lectures.

A good example of direct observation in a natural setting is Ya-Chung Chuang's study of democratization in Taiwan, *Democracy on Trial: Social-Movements and Cultural Politics in Post-Authoritarian Taiwan,* which uses ethnography to examine the interaction between individuals in a community and between individuals and institutions. To understand how individuals overcome complex problems, Chuang interviewed community and ethnic leaders, examined coalitions of community organizations, and interacted with residents in their communities. This method allowed Chuang to better understand a wide range of sociopolitical activities like community referendums for collective decision making or cultural walking tours

[7] William F. Whyte, *Street Corner Society: The Social Structure of an Italian Slum,* 3rd ed. (Chicago: University of Chicago Press, 1981); and Marc H. Ross, *Grass Roots in an African City: Political Behavior in Nairobi* (Cambridge, Mass.: MIT Press, 1975).

[8] Richard F. Fenno Jr., *Home Style: House Members in Their Districts* (Boston: Little, Brown, 1978).

[9] Ruth Horowitz, *Honor and the American Dream: Culture and Identity in a Chicano Community* (New Brunswick, N.J.: Rutgers University Press, 1983).

[10] Raphael Schlembach, "How do Radical Climate Movements Negotiate Their Environmental and Their Social Agendas? A Study of Debates within the Camp for Climate Action (UK)," *Critical Social Policy* 31, no. 2 (2011): 194–215.

[11] Joan E. McLean, "Campaign Strategy," in *Women and American Politics: New Questions, New Directions,* ed. Susan J. Carroll (Oxford, UK: Oxford University Press, 2003), 53–71.

in the community.[12] Chuang's personal experiences and firsthand observations in local communities gave him a vantage point that could not be acquired otherwise.

Direct observation can be carried out in many different ways, including as a participant or nonparticipant observer, in a structured or unstructured format, and as an overt or covert observer. In Chuang's study of Democracy in Taiwan he made observations as a participant observer—interacting with the people and institutions he was studying and participating in conversations and events as they unfolded. In **participant observation** the investigator is "both an actor and a spectator"—that is, a regular participant in the activities of the group being observed.[13] A researcher does not, however, have to become a full-fledged member of the. group to be a participant observer. Some mutually acceptable role or identity must be worked out. For example, Horowitz did not become a gang member when she studied Chicano youth in a Chicago neighborhood.[14] She hung around with gang members, but as a nonmember. She did not participate in fights and was able to decline when asked to conceal weapons for gang members. A nonparticipant observer does not participate in group activities or become a member of the group or community. For example, an investigator interested in hearings held by public departments of transportation or city council meetings could observe those proceedings without becoming a participant.

Most field studies involve participant observation. An investigator cannot be like the proverbial fly on the wall, observing a group of people for long periods of time. Usually he or she must assume a role or identity within the group under observation and participate in the activities of the group. In addition to interviewing influential Latinos in Boston, Carol Hardy-Fanta joined the community group Familias Latinas de Boston while conducting her research on Latina women and politics. As she pointed out, this strategy complemented her research interviews:

> Joining the community group Familias Latinas de Boston allowed me to gain an in-depth understanding of one community group over an extended period. Participating in formal, organized political activities such as manning the phone bank at the campaign office of a Latino candidate and attending political banquets, public forums, and conferences and workshops provided another means of observing how gender and culture interacted to stimulate—or suppress—political participation. I also joined protest marches and rallies and tracked down voter registration information in Spanish for a group at Mujeres Unidas en Acción. In addition, I learned much from informal interactions: at

[12] Chuang, Ya-Chung, *Democracy on Trial: Social Movements and Cultural Politics in Post-Authoritarian Taiwan* (Hong Kong: The Chinese University Press, 2013).

[13] Wedeen, "Reflections on Ethnographic Work in Political Science."

[14] Ruth Horowitz, "Remaining an Outsider: Membership as a Threat to Research Rapport," *Journal of Contemporary Ethnography* 14, no. 4 (1986): 409–30.

> groups on domestic violence, during lunch at Latino community centers, and during spontaneous conversations with Latinos from many countries and diverse backgrounds. As I talked to people in community settings and observed how they interacted politically, the political roles of Latina women and the gender differences in how politics is defined emerged. Thus, multiple observations were available to check what I was hearing in the interviews about how to stimulate Latino political participation, and how Latina women and Latino men act politically.[15]

Acceptance by the group is necessary for the investigator to benefit from the naturalness of the research setting. Negotiating an appropriate role for oneself within a group may be a challenging and evolving process. As Chicago gang researcher Ruth Horowitz pointed out, a researcher may not wish, or be able, to assume a role as a "member" of the observed group. Personal attributes (gender, age, ethnicity) of the researcher or ethical considerations (gang violence) may prevent this.[16] The role the researcher is able to establish also depends on the setting and the members of the group:

> I was able to negotiate multiple identities and relationships that were atypical of those generally found in the research setting, but that nonetheless allowed me to become sufficiently close to the setting members to do the research. By becoming aware of the nature, content, and consequences of these identities, I was able to use the appropriate identity to successfully collect different kinds of data and at the same time avoid some difficult situations that full participation as a member might have engendered.[17]

Participant observation is often used as one of several data collection methods in a single study. For example, Williamson, Skocpol, and Coggin used fieldwork observations and personal interviews along with an e-mail questionnaire of Massachusetts tea party activists. The authors used this observational data to supplement data from national surveys of the demographic and attitudinal characteristics of tea party activists and information on activism and ideology from local and regional tea party Web sites, among other sources.[18] J. C. Sharman used surveys and interviews as well as participant observation in his study of the adoption of anti-money-laundering policies in developing countries.[19]

[15] Carol Hardy-Fanta, *Latina Politics, Latino Politics: Gender, Culture, and Political Participation in Boston* (Philadelphia: Temple University Press, 1993), xiv.

[16] Horowitz, "Remaining an Outsider: Membership as a Threat to Research Rapport," 412.

[17] Ibid., 413.

[18] Vanessa Williamson, Theda Skocpol, and John Coggin, "The Tea Party and the Remaking of Republican Conservatism," *Perspectives on Politics* 9, no. 1 (2011): 25–43.

[19] J. C. Sharman, "Power and Discourse in Policy Diffusion: Anti–Money Laundering in Developing States," *International Studies Quarterly* 52, no. 3 (2008): 635–56.

Participant observation offers the advantages of a natural setting; the opportunity to observe people for lengthy periods so that interaction and changes in behavior may be studied; and a degree of accuracy or completeness that documents or recall data, such as that obtained in surveys, cannot provide. Observing a city council or school board meeting or a public hearing on the licensing of a locally undesirable land use will allow you to know and understand what happened at the event far better than reading official minutes or transcripts. However, this method has some noteworthy limitations as well.

The main problem with participant observation as a method of empirical research for political scientists is that many significant instances of political behavior are not accessible for observation. The privacy of the voter in the voting booth in the United States is legally protected, US Supreme Court conferences are not open to anyone but the justices themselves, and authoritarian regimes often design institutions that are purposely difficult to access and reject the idea of public disclosure. Occasionally, physical traces of these private behaviors become public—such as the Watergate tapes of Richard Nixon's conversations with his aides—and disclosures are made about some aspects of government decision making, such as congressional committee hearings and Supreme Court oral arguments. Typically, however, access is the major barrier to directly observing consequential political behavior.

Another disadvantage of participant observation is lack of control over the environment. A researcher may be unable to isolate individual factors and observe their effect on behavior. Participant observation is also limited by the small number of cases that are usually involved. For example, Fenno's research on "representatives' perceptions of their constituencies while they are actually in their constituencies"[20] was based on observations of only eighteen members or would-be members of Congress—too few for any sort of statistical analysis. He chose "analytical depth" over "analytical range"; in-depth observation of eighteen cases was the limit that Fenno thought he could manage intellectually, professionally, financially, and physically.[21] Whyte studied life in an Italian slum in *Street Corner Society* by observing one street corner gang in depth, although he did observe others less closely.[22] Because of the small numbers of cases, the representativeness of the results of participant observation has been questioned. But, as we stressed in our discussion of research designs (chapter 6), the number of cases deemed appropriate for a research topic depends on the purpose of the research. Understanding how people function in a particular community may be the knowledge that is desired, not whether the particular community is representative of some larger number of communities.

[20] Fenno, *Home Style: House Members in Their Districts*, xiii.

[21] Ibid., 255.

[22] Whyte, *Street Corner Society.*

Investigators using participant observation often depend on members of the group they are observing to serve as **informants**, persons who are willing to be interviewed about the activities and behavior of themselves and of the group to which they belong. An informant also helps the researcher interpret group behavior. A close relationship between the researcher and the informant may help the researcher gain access to other group members, not only because an informant may familiarize the researcher with community members and norms but also because the informant, through close association with the researcher, will be able to pass on information about the researcher to the community.[23] Some participant observation studies have one key informant; others have several. For example, Whyte relied on the leader of a street corner gang whom he called "Doc" as his key informant, while Fenno's eighteen representatives all could be considered informants.[24] In fact, interviewing members of the group being observed is an integral part of participation observation in most cases.

Although a valuable asset to researchers, informants may present problems. A researcher should not rely too much on one or a few informants, since they may give a biased view of a community. And if the informant is associated with one faction in a multifaction community or is a marginal member of the community (and thus more willing to associate with the researcher), the researcher's affiliation with the informant may inhibit rather than enhance access to the community.[25]

Structured and Unstructured Observation

In **structured observation**, the investigator looks for and systematically records the incidence of specific behaviors. The researcher will have decided, based on theory, the relevant behaviors before starting data collection. In **unstructured observation**, all behavior is considered relevant, at least at first, and recorded. Only later, upon reflection, will the investigator distinguish between important and trivial behavior. Open-ended, flexible observation is appropriate if the research purpose is one of description and exploration. For example, Fenno explained that he began unstructured data collection in order to crystalize his thoughts on what was important to study about members of Congress in their districts. As Fenno explained, his visits with representatives in their districts

> were totally open-ended and exploratory. I tried to observe and inquire into anything and everything these members did. I worried about whatever they worried about. Rather than assume that I already knew

[23] Jennie-Keith Ross and Marc Howard Ross, "Participant Observation in Political Research," *Political Methodology* 1 (1974): 70.

[24] Whyte, *Street Corner Society;* Fenno, *Home Style: House Members in Their Districts.*

[25] Ross and Ross, "Participant Observation in Political Research."

> what was interesting, I remained prepared to find interesting questions emerging in the course of the experience. The same with data. The research method was largely one of soaking and poking or just hanging around.[26]

In these kinds of field studies, researchers do not start out with particular hypotheses that they want to test. They often do not know enough about what they plan to observe to establish lists and specific categories of behaviors to look for and record systematically. The purpose of the research is to discover what these might be.

Some political scientists have used observation as a preliminary research method.[27] For example, James A. Robinson's work in Congress provided firsthand information for his studies of the House Rules Committee and of the role of Congress in making foreign policy.[28] Ralph K. Huitt's service on Lyndon B. Johnson's Senate majority leader staff gave Huitt inside access to information for his study of Democratic Party leadership in the Senate.[29] And David W. Minar served as a school board member and used his experience to develop questionnaires for his comparative study of several school districts in the Chicago area.[30] As mentioned earlier, Carroll and Liebowitz suggested observing women's activities in order to identify behaviors with political effect that have not previously been included in measures of political activity; subsequent surveys could then include questions that ask about such behaviors.[31]

Unstructured participant observation also has been criticized as invalid and biased. A researcher may selectively perceive behaviors, noting some while ignoring others. The interpretation of behaviors may reflect the personality and culture of the observer rather than the meaning attributed to them by the observed themselves. Moreover, the presence of the observer may alter the behavior of the observed, no matter how skillfully the observer attempts to become accepted as a nonthreatening part of the community.

Fieldworkers attempt to minimize these possible threats to data validity by immersing themselves in the culture they are observing and by taking copious notes on

[26] Ibid., xiv.

[27] Ibid., 65–66.

[28] James A. Robinson, *The House Rules Committee* (Indianapolis, Ind.: Bobbs-Merrill, 1963); and James A. Robinson, *Congress and Foreign Policy-Making: A Study in Legislative Influence and Initiative* (Homewood, Ill.: Dorsey Press, 1962). Also, extensive firsthand observations of Congress are reported in many of the articles in Raymond E. Wolfinger, ed., *Readings on Congress* (Englewood Cliffs, N.J.: Prentice Hall, 1971).

[29] Ralph K. Huitt, "Democratic Party Leadership in the Senate," *American Political Science Review* 55, no. 2 (1961): 333–44.

[30] David W. Minar, "The Community Basis of Conflict in School System Politics," in *The New Urbanization*, ed. Scott Greer et al. (New York: St. Martin's Press, 1968), 246–63.

[31] Carroll and Liebowitz, "Introduction: New Challenges, New Questions, New Directions."

everything going on around them, no matter how seemingly trivial. Events without apparent meaning at the time of observation may become important and revealing upon later reflection. Of course, copious note-taking leads to what is known as a "high dross rate"; much of what is recorded is not relevant to the research problem or question as it is finally formulated. It may be painful for the investigator to discard so much of the material that was carefully recorded, but it is standard practice with this method.

Another way to obtain more valid data is to allow the observed to read and comment on what the investigator has written and point out events and behaviors that may have been misinterpreted. This check on observations may be of limited or no value if the person being observed cannot read or if the written material is aimed at persons well versed in the researcher's discipline and therefore is over the head of the observed.

Researchers' observations may be compromised if the researchers begin to overidentify with their subjects or informants. "Going native," as this phenomenon is known, may lead researchers to paint a more complimentary picture of the observed than is warranted. Researchers combat this problem by returning to their own culture to analyze their data and by asking colleagues or others to comment on their findings.

Covert and Overt Observation

Another choice in direct observation is between overt or covert observations. In **overt observation**, those being observed are aware of the investigator's presence and intentions. In **covert observation**, the investigator's presence is hidden or undisclosed, and his or her intentions are disguised. For example, covert observation was used in a study to measure what percentage of people washed their hands after using the restroom.[32] The advantage of covert observation is that the researcher may be better able to observe unrestrained behavior. If people are aware that someone is watching and recording observations they may behave differently than they normally would. Hence, by concealing observation a researcher may be able to make more valid observations. Research involving covert observation of public behavior of private individuals is not likely to raise ethical issues as long as individuals are not or cannot be identified and disclosure of individuals' behavior would not place them at risk. Note that elected or appointed public officials are not shielded by these limitations. Ethical standards and their application or enforcement have changed, and it is likely that many earlier examples of participant observation research, especially those involving covert observation, would not receive

[32] Paul B. Allwood, "Handwashing among Public Restroom Users at the Minnesota State Fair." Accessed January 21, 2015. Available at http://www.health.state.mn.us/handhygiene/stats/fairstudy.pdf

approval from human subject review boards today. For example, social scientists Mary Henle and Marian B. Hubble once hid under beds in students' rooms to study student conversations.[33]

Note-Taking as Data Collection

A demanding, yet essential, aspect of field study is note-taking. Notes can be divided into three types: mental notes, jotted notes, and field notes. Mental note-taking involves orienting one's consciousness to the task of remembering things one has observed, such as "who and how many were there, the physical character of the place, who said what to whom, who moved about in what way, and a general characterization of an order of events."[34] Because mental notes may fade rapidly, researchers use jotted notes to preserve them. Jotted notes consist of short phrases and keywords that will activate a researcher's memory later when the full field notes are written down. Researchers may be able to use tape recorders if they have the permission of those being observed.

Taped conversations do not constitute "full" field notes, which should include a running description of conversations and events. For this aspect of field notes, John Lofland advised that researchers should be factual and concrete, avoid making inferences, and use participants' descriptive and interpretative terms. Full field notes should include material previously forgotten and subsequently recalled. Lofland suggested that researchers distinguish between verbal material that is exact recall, paraphrased or close recall, and reasonable recall.[35]

Field notes should also include a researcher's analytic ideas and inferences, personal impressions and feelings, and notes for further information.[36] Because events and emotional states in a researcher's life may affect observation, they should be recorded. Notes for further information provide guidance for future observation—to fill in gaps in observations, call attention to things that may happen, or test out emerging analytic themes.

Full field notes should be legible and should be reviewed periodically, since the passage of time may present past observations in a new light to the researcher or reveal a pattern worthy of attention in a series of disjointed events. Creating and reviewing field notes is an important part of the observational method. Consequently, a fieldworker should expect to spend as much time on field notes as he

[33] Mary Henle and Marian B. Hubble, "'Egocentricity' in Adult Conversation," *Journal of Social Psychology* 9, no. 2 (1938): 227–34..

[34] John Lofland, *Analyzing Social Settings: A Guide to Qualitative Observation and Analysis* (Belmont, Calif.: Wadsworth, 1971), 102–03.

[35] Ibid., 105.

[36] Ibid., 106–7.

or she spends on observation in the field. Fortunately, computerized text analysis programs exist to help analyze field notes and interviews.

Note-taking is especially important because the recorded observations in notes constitute the data in fieldwork. The narratives that researchers develop and the conclusions they make come directly from the field notes. The richness of detail in the notes leads to a fully developed, published product that relays to the reader a sense of what the researcher experienced. Note-taking is therefore the lifeblood of field research, as it records the observations and anecdotes that eventually find their way to the printed page.

Direct Observation in a Laboratory Setting

Observation in a laboratory setting gives a researcher the advantage of having control over the environment of the observed. For example, Mendelberg, Karpowitz, and Oliphant executed an observational design in a laboratory setting to test how decision rules like majority rule and consensus affect conversation dynamics between men and women. The project was designed to give better leverage in understanding the mechanism at work in critical-mass theory of how the number of women in a legislature affects the culture of interaction between men and women. In brief, the theory is that when women make up less than 15 percent of a legislative body, men dominate the culture and conversation, but when women exceed 35 percent, the culture begins to change and women take on a stronger voice in conversation, with more of a role. The authors theorized that decision rules, majority rule, or consensus also might affect the relationship between men and women in conversation. To test this hypothesis, the authors observed small groups in a laboratory setting that included both men and women as they discussed and decided on the best way to address principles of income distribution. The authors concluded that the decision rule did have an effect—women participated more and were interrupted less frequently when the decision rule was consensus rather than majority rule. The laboratory setting was crucial to this design because it allowed the authors to control important aspects of the interaction including the decision rule, the size of the groups, the ratio of men to women, and much more while observing every word.[37] Thus, these researchers were able to use a more rigorous experimental design than would have been possible in a natural, uncontrolled setting. In addition to the ability to control the setting, observation may be easier and more convenient to record and preserve, since one-way windows, video cameras, and other observational aids are more readily available in a laboratory.

[37] Tali Mendelberg, Christopher F. Karpowitz, and J. Baxter Oliphant, "Gender Inequality in Deliberation: Unpacking the Black Box of Interaction," *Perspectives on Politics* 12, no. 1 (2014).

A disadvantage of laboratory observation is that subjects usually know they are being observed and therefore may alter their behavior, raising questions about the validity of the data collected. The use of aids that allow the observer to be physically removed from the setting and laboratories that are designed to be as inviting and natural as possible may lead subjects to behave more naturally and less self-consciously.

An example of an attempt to create a natural-looking laboratory setting may be found in Stanley Milgram and R. Lance Shotland's book *Television and Antisocial Behavior.*[38] These researchers were interested in the effect of television programming on adult behavior, specifically in the ability of television drama to stimulate antisocial acts such as theft. They devised four versions of a program called *Medical Center,* each with a different plot, and showed different versions to four different audiences. Some of the versions showed a character stealing money, and those versions differed in whether the person was punished for the theft or not. The participants in the study were then asked to go to a particular office at a particular time to pick up a free transistor radio, their payment for participating in the research study. When they arrived in the office (the laboratory), they encountered a sign that said the radios were all gone. The researchers were interested in how people would react and specifically in whether they would imitate any of the behaviors in the versions of *Medical Center* that they had seen (such as the theft of money from see-through plastic collection dishes). Their behavior was observed covertly via a one-way mirror. Once the subjects left the office, they were directed to another location where they were, in fact, given the promised radio. (This experiment, reported in 1973, raises some serious ethical issues about deceiving research subjects and causing them harm.)

Firsthand, Indirect Observation

While most political science research using firsthand observation uses a direct observation method, some researchers rely on indirect observation instead. Indirect observation, the observation of physical traces of behavior, is essentially detective work.[39] Inferences based on physical traces can be drawn about people and their behavior. An unobtrusive research method, indirect observation is nonreactive: subjects do not change their behavior because they do not know they are being studied.

[38] Stanley Milgram and R. Lance Shotland, *Television and Antisocial Behavior: Field Experiments* (New York: Academic Press, 1973).

[39] Webb, Campbell, and Schwartz, *Nonreactive Measures in the Social Sciences,* 4.

Physical Trace Measures

Researchers use two methods of measurement when undertaking indirect observation. An **erosion measure** is created by selective wear on some material. For example, campus planners at one university observed paths worn in grassy areas and then rerouted paved walkways to correspond to the most heavily trafficked routes. Other examples of natural erosion measures include wear on library books; wear and tear on selected articles within volumes; and depletion of items in stores, such as by sales of newspapers.

The second measurement of indirect observation is the **accretion measure**, which measures a phenomenon as manifested through the deposition and accumulation of materials. Archaeologists and geologists commonly use accretion measures in their research by measuring, mapping, and analyzing accretion of materials. Other professions find them useful as well. Eugene Webb and his colleagues reported a study in which mechanics in an automotive service department recorded radio dial settings to estimate radio station popularity.[40] This information was then used to select radio stations to carry the dealer's advertising. The popularity of television programs could be measured by recording the drop in water level in community water-storage systems while commercials are aired, since viewers tend to use the toilet only during commercials when watching very popular shows. Or the reverse could be explored to test the popular wisdom that commercials shown during the Super Bowl are more popular than the game itself. Similarly, declines in telephone usage could indicate television program popularity. The presence of fingerprints and nose prints on glass display cases may indicate interest as well as reveal information about the size and age of those attracted to the display. The effectiveness of various antilitter policies and conservation programs could also be measured using physical trace evidence, and the amount and content of graffiti may represent an interesting measurement of the beliefs, attitudes, and mood of a population.

One of the best-known examples of the use of accretion measures is W. L. Rathje's study of people's garbage.[41] He studied people's behavior based on what they discarded in their trash cans. One project involved investigating whether poor people wasted more food than those better off; they did not.

Indirect observation typically raises fewer ethical issues than direct observation because the measures of individual behavior are taken after the individuals have left the scene, thus ensuring anonymity in most cases. However, Rathje's studies of garbage raised ethical concerns because some discarded items (such as letters and bills) identified the source of the garbage. Although a court ruled in Rathje's

[40] Ibid., 10–11.

[41] See discussion of Rathje's work in ibid., 15–17.

favor by declaring that when people discard their garbage, they have no further legal interest in it, one might consider sorting through a person's garbage to be an invasion of privacy. In a study in which data on households were collected, consent forms were obtained, codes were used to link household information to garbage data, and then the codes were destroyed. Rathje's assistants in another garbage study were instructed not to examine any written material closely.

It is also possible that garbage may contain evidence of criminal wrongdoing. Twice during Rathje's research, body parts were discovered, although not in the bags collected as part of the study. Rathje took the position that evidence of victimless crimes should be ignored but evidence of serious crimes should be reported. Of course, the publicity surrounding Rathje's garbage study may have deterred disposal of such evidence. This raises the problem of reactivity: To what extent might people change their garbage-disposing habits if they know there is a small chance that what they throw away will be examined?

This example also illustrates the possibility that indirect observation of physical traces of behavior may border on direct observation of subjects if the observation of physical traces quickly follows their creation. In some situations, extra measures may have to be taken to preserve the anonymity of subjects.

Another good example of the use of accretion measures is Kurt Lang and Gladys Engel Lang's study of the MacArthur Day parade in Chicago in 1951.[42] Gen. Douglas MacArthur and President Harry S. Truman were locked in an important political struggle at the time, and the Langs wanted to find out how much interest there was in the parade. They used data on mass-transit passenger fares, hotel reservations, retail store and street vendor sales, parking lot usage, and the volume of ticker tape on the streets to measure the size of the crowd attracted by MacArthur's appearance.

Validity Problems with Indirect Observation

Although physical trace measures generally are not subject to reactivity to the degree that participant observation and survey research are, threats to the validity of these measures do exist. Also, erosion and accretion measures may be biased. For example, certain traces are more likely to survive because the materials are more durable. Thus, physical traces may provide a selective, rather than complete, picture of the past. Differential wear patterns may be due not to variation in use but to differences in material. Researchers studying garbage must be careful not to infer that garbage reflects all that is used or consumed. Someone who owns a garbage disposal, for example, generally discards less garbage than someone who does not.

[42] Kurt Lang and Gladys Engel Lang, *Politics and Television* (Chicago: Quadrangle Books, 1968).

Researchers should exercise caution in linking changes in physical traces to particular causes. Other factors may account for variation in the measures. Webb and his colleagues suggested that several physical trace measures be used simultaneously or that alternative data collection methods be used to supplement physical trace measures.[43] For example, physical trace measures of the use of recreational facilities, such as which trash cans in a park fill up the fastest, could be supplemented with questionnaires completed by park visitors on facility usage.

Caution should also be used in making inferences about the behavior that caused the physical traces. For example, wear around a particular museum exhibit could indicate either the number of people viewing the exhibit or the amount of time people spent near the exhibit shuffling their feet. Direct observation could determine the answer, but in cases where the physical trace measures occurred in the past, this solution is not possible.

Examples of the use of indirect observation in political science research are not numerous. Nevertheless, this method has been used profitably, and you may be able to think of cases where it would be appropriate. For example, you could assess the popularity of candidates by determining the number of yard signs appearing in a community. Or you could estimate the number of visitors and level of office activity of elected representatives by noting carpet wear in office entryways. Although this would not be as precise as counting visitors, it would allow you to avoid posting observers or questioning office staff.

Indirect observation, when used ingeniously, can be a low-cost research method free from many of the ethical issues that surround direct observation. Let us now turn to a consideration of some of the ethical issues that develop in the course of fieldwork and in simple, nonexperimental laboratory observations.

Ethical Issues in Observation

Ethical dilemmas arise primarily when there is a potential for harm to the observed. The potential for serious harm to subjects in most observational studies is quite low. Observation generally does not entail investigation of highly sensitive, personal, or illegal behavior, because people are reluctant to be observed in those circumstances and would not give their informed consent. Nor do fieldwork and simple laboratory observation typically involve experimental manipulations of subjects and exposure to risky experimental treatments. Nonetheless, harm or risks to the observed may result from observation. They include (1) negative repercussions from associating with the researcher because of the researcher's sponsors, nationality, or outsider status;

[43] See Webb, Campbell, and Schwartz, *Nonreactive Measures in the Social Sciences*, 27–32.

(2) invasion of privacy; (3) stress during the research interaction; and (4) disclosure of behavior or information to the researcher resulting in harm to the observed during or after the study. Each of these possibilities is considered here in turn.

In some fieldwork situations, contact with outsiders may be viewed as undesirable behavior by an informant's peers. Cooperation with a researcher may violate community norms. For example, a researcher who studies a group known to shun contact with outsiders exposes informants to the risk of being censured by their group.

Social scientists from the United States have encountered difficulty in conducting research in countries that have hostile relations with the United States.[44] Informants and researchers may be accused of being spies, and informants may be exposed to harm for appearing to sympathize with "the enemy." Harm may result even if hostile relations develop after the research has been conducted. Military, Central Intelligence Agency, or other government sponsorship of research may particularly endanger the observed.

A second source of harm to the observed results from the invasion of privacy that observation may entail. Even though a researcher may have permission to observe, the role of observer may not always be remembered by the observed. In fact, as a researcher gains rapport, there is a greater chance that informants may view the researcher as a friend and reveal to him or her something that could prove to be damaging. A researcher does not always warn, "Remember, you're being observed!" Furthermore, if a researcher is being treated as a friend, such a warning may damage rapport. Researchers must consider how they will use the information gathered from subjects. They must judge whether use in a publication will constitute a betrayal of confidence.[45]

Much of the harm to subjects in fieldwork occurs as a result of publication. They may be upset at the way they are portrayed, subjected to unwanted publicity, or depicted in a way that embarrasses the larger group to which they belong. Carelessness in publication may result in the violation of promises of confidentiality and anonymity. And value-laden terminology may offend those being described:[46]

In accordance with federal regulations, universities and other research organizations require faculty and students to submit research proposals involving human subjects for review by an **institutional review board** (often called a human subject review board). There may be some variation in practice concerning unfunded

[44] See Myron Glazer, *The Research Adventure: Promise and Problems of Field Work* (New York: Random House, 1972), 25–48, 97–124.

[45] See Fenno, *Home Style: House Members in Their Districts,* 272.

[46] For a discussion and examples of value-laden terminology in published reports of participant observers, see ibid.

research, but the proper course of action is to contact your institution's research office for information regarding the review policy on human subjects. There are three levels of review: some research may be exempt, some may require only expedited review, and some research will be subject to full board review. Even if your research project seems to fit one of the categories of research exempt from review, you must request and be granted an exemption.[47]

Three ethical principles—respect for persons, beneficence, and justice—form the foundation for assessing the ethical dimensions of research involving human subjects. These principles were identified in the *Belmont Report,* a report of the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research.[48] The principle concerning respect for persons asserts that individuals should be treated as autonomous agents and that persons with diminished capacity are entitled to protection. *Beneficence* refers to protecting people from harm as

[47] Exemption categories are as follows: "1. Research conducted in established or commonly accepted educational settings, involving normal educational practices, such as (a) research on regular and special education instructional strategies or (b) research on the effectiveness of or the comparison among instructional techniques, curricula, or classroom management methods. 2. Research involving the use of educational tests (cognitive, diagnostic, aptitude, achievement), survey procedures, interview procedures, or observation of public behavior, unless (a) information obtained is recorded in such a manner that human subjects can be identified, directly or through identifiers linked to the subjects, AND (b) any disclosure of the human subjects' responses outside the research could reasonably place the subjects at risk of criminal or civil liability or be damaging to the subjects' financial standing, employability, or reputation. 3. Research involving the use of education tests, survey procedures, interview procedures, or observation of public behavior that is not exempt under category 2, if (a) the human subjects are elected or appointed public officials or candidates for public office or (b) federal statute(s) requires without exception that the confidentiality of the personally identifiable information will be maintained throughout the research and thereafter. 4. Research involving the collection or study of existing data, documents, records, pathological specimens, or diagnostic specimens, if these sources are publicly available or if the information is recorded by the investigator in such a manner that subjects cannot be identified directly or through identifiers linked to the subjects. 5. Research and demonstration projects that are conducted by or subject to the approval of department or agency heads and that are designed to study, evaluate, or otherwise examine (a) public benefit or service programs, (b) procedures for obtaining benefits or services under those programs, (c) possible changes in or alternatives to those programs or procedures, or (d) possible changes in methods or levels of payment for benefits or services under those programs. 6. Taste and food quality evaluation and consumer acceptance studies, (a) if wholesome foods without additives are consumed or (b) if a food is consumed that contains a food ingredient at or below the level and for a use found to be safe, or agricultural chemical or environmental contaminant at or below the level found to be safe, by the Food and Drug Administration or approved by the Environmental Protection Agency or the Food Safety and Inspection Service of the U.S. Department of Agriculture." From United States Office of the Federal Register, *Code of Federal Regulations: Title 45, Public Welfare; Part 46, Protection of Human Subjects* (Washington, D.C.: US Government Printing Office, 1977), Part 46.101(b). These exemptions do not apply to research involving prisoners, fetuses, pregnant women, or human in vitro fertilization. Exemption 2 does not apply to children except for research involving observations of public behavior when the investigator does not participate in the activities being observed.

[48] National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, *The Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research* (Washington, D.C.: US Government Printing Office, 1979). Available at http://ohsr.od.nih.gov/guidelines/belmont.html

well as making efforts to secure their well-being. The principle of justice requires researchers to consider the distribution of the benefits and burdens of research.

The principle of respect for persons requires that subjects be given the opportunity to choose what shall or shall not happen to them. **Informed consent** means that subjects are to be given information about the research, including the research procedure, its purposes, risks, and anticipated benefits; alternative procedures (where therapy is involved); how subjects are selected; and the person responsible for the research. In addition, the subject is to be given a statement offering him or her the opportunity to ask questions and to withdraw from the research at any time. This information and statement should be conveyed in a manner that is comprehensible to the subject, and the consent of the subject must be voluntary.

An assessment of risks and benefits relates directly to the beneficence principle by helping to determine whether risks to subjects are justified and by providing information useful to subjects for their informed consent. The justice principle is often associated with the selection of subjects insofar as some populations may be more likely to be targeted for study; one example is prison populations, particularly in the past.

Conclusion

Firsthand observation is an important research method for political scientists. Observational studies may be direct or indirect. Indirect observation is less common but has the advantage of being a nonreactive research method. Direct observation of people by social scientists has produced numerous studies that have enhanced knowledge and understanding of human beings and their behavior. Fieldwork—direct observation by a participant observer in a natural setting—is the best-known variety of direct observation, although direct observation may take place in a laboratory setting. Observation tends to produce data that are qualitative rather than quantitative. Because the researcher is the measuring device, this method is subject to particular questions about researcher bias and data validity. Since there is an evolving relationship between the observer and the observed, participant observation is a demanding and often unpredictable research endeavor. Part of the demanding nature of fieldwork stems from the difficult ethical dilemmas it raises.

As a student you may find yourself in the position of an observer, but it is more likely that you will be a consumer and evaluator of observational research. In this position you should base your evaluation on many considerations: Does it appear that the researcher influenced the behavior of the observed or was biased in his or her observation? How many informants were used, a few or only one? Does it appear likely that the observed could have withheld significant behavior of interest to the researcher? Are generalizations from the study limited because observation was made in a laboratory setting or because of the small number of cases observed? Were any ethical issues raised by the research? Could they have been avoided? What would you have done in a similar situation? Asking these questions will help you evaluate the validity and ethics of observational research.

TERMS INTRODUCED

Accretion measure. Measure of phenomena through indirect observation of the accumulation of materials.

Covert observation. Observation in which the observer's presence or purpose is kept secret from those being observed.

Direct observation. Actual observation of behavior.

Erosion measure. Measure of phenomena through indirect observation of selective wear of some material.

Ethnography. A type of field study in which the researcher is deeply immersed in the place and lives of the people being studied.

Field study. Open-ended and wide-ranging (rather than structured) observation in a natural setting.

Indirect observation. Observation of physical traces of behavior.

Informants. Persons who are willing to be interviewed about the activities and behavior of themselves and of the group to which they belong. An informant also helps the researcher engaged in participant observation to interpret group behavior.

Informed consent. Procedures that inform potential research subjects about the proposed research in which they are being asked to participate; the principle that researchers must obtain the freely given consent of human subjects before they participate in a research project.

Institutional review board. Panel to which researchers must submit descriptions of proposed research involving human subjects for the purpose of ethics review.

Overt observation. Observation in which those being observed are informed of the observer's presence and purpose.

Participant observation. Observation in which the observer becomes a regular participant in the activities of those being observed.

Primary data. Data recorded and used by the researcher who is making the observations.

Reactivity. Effect of data collection or measurement on the phenomenon being measured.

Secondary data. Data used by a researcher that were not personally collected by that researcher.

Structured observation. Systematic observation and recording of the incidence of specific behaviors.

Unstructured observation. Observation in which all behavior and activities are recorded.

SUGGESTED READINGS

Fenno, Richard F., Jr. *Home Style: House Members in Their Districts.* Boston: Little, Brown, 1978. See esp. the introduction and appendix, "Notes on Method: Participant Observation."

Jaggar, Alison M. *Just Methods: An Interdisciplinary Reader.* Boulder, Colo.: Paradigm, 2008.

Piccolo, Francesco Lo, and Huw Thomas, eds. *Ethics and Planning Research.* Burlington, Vt.: Ashgate, 2009.

Reason, Peter, and Hilary Bradbury, eds. *The SAGE Handbook of Action Research: Participative Inquiry and Practice.* 2nd ed. Thousand Oaks, Calif.: Sage, 2008.

Sieber, Joan E., ed. *The Ethics of Social Research: Fieldwork, Regulation, and Publication.* New York: Springer-Verlag, 1982.

Sieber, Joan E., and Martin B. Tolich. *Planning Ethically Responsible Research: A Guide for Students and Internal Review Boards.* 2nd ed. Applied Social Research Methods Series vol. 31. Newbury Park, Calif.: Sage, 2013.

Smyth, Marie, and Emma Williamson, eds. *Researchers and Their "Subjects": Ethics, Power, Knowledge, and Consent.* Bristol, UK: Policy Press, 2004.

Wedeen, Lisa. "Reflections on Ethnographic Work in Political Science." *Annual Review of Political Science* 13, no. 1 (2010): 255–72.

Williamson, Vanessa, Theda Skocpol, and John Coggin. "The Tea Party and the Remaking of Republican Conservatism." *Perspectives on Politics* 9, no. 1 (2011): 25–43.

CHAPTER 9

Document Analysis:

Using the Written Record

CHAPTER OBJECTIVES

9.1 Explain the role and procedures of content analysis.

9.2 Identify different types of written records.

9.3 Contrast the advantages and disadvantages of the written record.

IN THIS CHAPTER WE DESCRIBE DOCUMENT ANALYSIS—how empirical observations can be made using the **written record**, which is composed of documents, reports, statistics, manuscripts, and other written, oral, or visual materials.

Some research questions can be answered by examining records or data collected by others. Some of these materials have been purposefully collected over time by organizations for the purpose of study, like records collected by a presidential library, while others are records that have accumulated after serving another purpose, like newspaper articles in an archive. These records and data have become increasingly available in recent years as governments, businesses, and academics digitize records, making many available on the Internet. For some questions, other data collection methods such as interviewing and firsthand observation are of limited utility to researchers interested in large-scale collective behavior (such as civil unrest and the budget allocations of national governments) or in phenomena that are distant in time (Supreme Court decisions during the Civil War) or space (defense spending by different countries).

The political phenomena that have been observed through written records are many and varied—for example, judicial decisions concerning the free exercise of religion, voter turnout rates in gubernatorial elections, the change over time in Russian military expenditures, and the incidence of political corruption in the People's Republic of China.[1] Of the examples of political science research described in chapter 1 and referred to throughout this book, Lane Kenworthy and Jonas Pontusson's and Jacob S. Harker and Paul Peterson's studies of income inequality; Thomas Holbrook and Brianne Heidbreder's study of voter turnout rates; Wesley T. Milner, Steven C. Poe, and David Leblang's investigation of governments' violation of human rights; Richard L. Hall and Kristina C. Miler's study of congressional oversight activity; Jeffrey A. Segal and Albert D. Cover's investigation of the ideology of Supreme Court justices; and several of the studies of the impact of negative campaign advertisements all depended on written records for the measurement of important political concepts.[2] Not all portions of the written record are equally

[1] Frank Way and Barbara J. Burt, "Religious Marginality and the Free Exercise Clause," *American Political Science Review* 77, no. 3 (1983): 652–65; Samuel C. Patterson and Gregory A. Caldeira, "Getting Out the Vote: Participation in Gubernatorial Elections," *American Political Science Review* 77, no. 3 (1983): 675–89; William Zimmerman and Glenn Palmer, "Words and Deeds in Soviet Foreign Policy: The Case of Soviet Military Expenditures," *American Political Science Review* 77, no. 2 (1983): 358–67; and Alan P. L. Liu, "The Politics of Corruption in the People's Republic of China," *American Political Science Review* 77, no. 3 (1983): 602–23.

[2] Lane Kenworthy and Jonas Pontusson, "Rising Inequality and the Politics of Redistribution in Affluent Countries," *Perspectives on Politics* 3, no. 3 (2005): 449–71, available at http://www.u.arizona.edu/~lkenwor/pop2005.pdf; Jacob S. Hacker and Paul Pierson, "Winner-Take-All Politics: Public Policy, Political Organization, and the Precipitous Rise of Top Incomes in the United States," *Politics & Society* 38, no. 2 (2010): 152–204; Thomas Holbrook and Brianne Heidbreder, "Does Measurement Matter? The Case of VAP and VEP in Models of Voter Turnout in the United States," *State Politics & Policy Quarterly* 10, no. 2 (2010): 159–81; Wesley T. Milner, Stephen C. Poe, and David Leblang, "Security Rights, Subsistence Rights, and Liberties: A Theoretical Survey of the Empirical Landscape," *Human Rights Quarterly* 21, no. 2 (1999): 403–43; Richard C. Hall and Kristina Miler, "What Happens after the Alarm? Interest Group Subsidies to Legislative Overseers," *Journal of Politics* 70, no. 4 (2008): 990–1005; Jeffrey A. Segal and Albert D. Cover, "Ideological Values and the Votes of U.S. Supreme Court Justices," *American Political Science Review* 83, no. 2 (1989): 557–65, available at http://www.uic.edu/classes/pols/pols200mm/Segal89.pdf; Stephen D. Ansolabehere, Shanto Iyengar, and Adam Simon, "Replicating Experiments Using Aggregate and Survey Data: The Case of Negative Advertising and Turnout," *American Political Science Review* 93, no. 4 (1999): 901–10; Stephen D. Ansolabehere, Shanto Iyengar, Adam Simon, and Nicholas Valentino, "Does Attack Advertising Demobilize the Electorate?" *American Political Science Review* 88, no. 4 (1994): 829–38, available at http://weber.ucsd.edu/~tkousser/Ansolabehere.pdf; Martin P. Wattenberg and Craig Leonard Brians, "Negative Campaign Advertising: Demobilizer or Mobilizer?" *American Political Science Review* 93, no. 4 (1999): 891, available at http://weber.ucsd.edu/~tkousser/Wattenberg.pdf; and Richard R. Lau and Ivy Brown Rovner, "Negative Campaigning," *Annual Review of Political Science* 12 (2009): 285–306.

useful to political scientists. Hence, we discuss the major components of the written record of interest to political scientists and how researchers use those components to measure significant political phenomena.

Content Analysis

Those who rely on the written record often extract excerpts, quotations, or examples from documents to support an observation or relationship. We can think of document analysis, much like other forms of analysis, as taking both a qualitative and a quantitative form. Qualitative document analysis relies on describing examples from records to explain political phenomena. Consider a research project on the ideology of Supreme Court justices. A researcher may use a qualitative approach that identifies patterns across the writings of different justices by analyzing quotations from their written opinions on various questions about the role of government or social and economic issues that come before the Court. Alternatively, a researcher might take an approach that involves applying systematic measurement to qualitative sources of information to measure justices' ideology to create quantitative data. This use of the written record via systematic coding and classification of its contents is an example of **content analysis**. A researcher uses content analysis by taking "a verbal, nonquantitative document and transform it into quantitative data."[3] A researcher "first constructs a set of mutually exclusive and exhaustive categories that can be used to analyze documents, and then records the frequency with which each of these categories is observed in the documents studied."[4] This is how Segal and Cover analyzed newspaper editorials to produce a quantitative measure of the Supreme Court justices' political ideologies.[5] Segal and Cover were able to first use their content analysis to create ideological scores for each justice and then use those scores to predict voting behavior in Supreme Court cases. In this section, we focus primarily on quantitative content analysis for use in statistical analyses, but remember that a qualitative approach to documents can be just as useful, if not more so, depending on the purpose of a research project.

Content Analysis Procedures

The first step in content analysis is to decide what materials to include in the analysis. This selection, of course, is guided by the topic, theory, existing research, etc. If a researcher is interested in the political values of candidates for public office, position papers and campaign speeches might be suitable. Or if a researcher is

[3] Kenneth D. Bailey, *Methods of Social Research*, 2nd ed. (New York: Free Press, 1982), 312–13.

[4] Ibid.

[5] Segal and Cover, "Ideological Values and the Votes of U.S. Supreme Court Justices."

interested in what liberals are currently thinking about the role of government in society, liberal opinion magazines or blogs might be used. Krippendorff referred to these materials as the "sampling units" and defined them as "units that are distinguished for selective inclusion in an analysis."[6] The list of materials germane to the researcher's subject thus makes up a "sampling frame." Once the appropriate sampling frame has been selected, then any of the possible types of samples described in chapter 7—random, systematic, stratified, cluster, and nonprobability—could be used. Of course, it may be the case that the sampling frame corresponds to the population and that all units of the population will be studied. For example, you might have all State of the Union addresses and wish to analyze all of them.

The second task in any content analysis is to define the "recording or coding units"—that is, "the units that are distinguished for separate description, transcription, recoding, or coding."[7] For example, from a given document, news item, video clip, or other material, the researcher may want to code (1) each word or sentence fragment, (2) each character or actor, (3) each sentence, (4) each paragraph, or (5) each item in its entirety. The choice of the recording or coding unit depends on the categories of content that are going to be measured. In choosing the recording unit, the researcher usually considers the correspondence between the unit and the content categories (stories may be more appropriate than words to determining whether crime is a topic of concern, whereas individual words or sentences rather than larger units may be more appropriate to measuring the traits of political candidates). Generally, if the recording unit is too small, it will be unlikely to possess any of the content categories. If the recording unit is too large, however, it will be difficult to measure the single category of a content variable that it possesses (in other words, the case will possess multiple values of a given content variable). For example, a paragraph or a story may contain both positive and negative evaluations of a candidate. The selection of the appropriate recording unit is often a matter of trial and error, adjustment, and compromise in the pursuit of measures that capture the content of the material being coded.

The third task, therefore, is to choose categories of content that are going to be measured. These categories are the variables you want to focus on in your study. This process is in many respects the most important part of any content analysis, because the researcher must measure the content in such a way that it relates to the research topic and must define this content so that the measures of it are both valid and reliable. So, for example, researchers studying the prevalence of crime in the news might take a sample of the front pages of newspapers or half-hour nightly news programs and measure the amount of content that either deals with crime or does not.

[6] Klaus Krippendorff, *Content Analysis: An Introduction to Its Methodology*, 2nd ed. (Thousand Oaks, Calif.: Sage, 2004), 98.

[7] Ibid., 99.

Content analysis can be greatly improved in quality and efficiency by taking advantage of computer software designed to identify and code qualitative material. In its most basic form researchers can identify keywords and phrases that define how a software package will identify relevant words, phrases, or sections of written documents—this is similar to a keyword search you might use with an Internet search engine or your library's electronic catalog (see our example below). Once keywords or passages have been identified the software can be used to automatically code documents for use in quantitative analysis or allow the researcher to code identified passages manually. Some software can use examples of manually coded material and apply the examples to code additional material. Once coded, software can be used to analyze data and create graphical representations of relationships using charts, tables, and figures.[8]

The validity of a content analysis can usually be enhanced with a precise explanation of the procedures followed and content categories used. Usually the best way to demonstrate the reliability of content analysis measures is to show intercoder reliability. **Intercoder reliability** simply means that two or more analysts, using the same procedures and definitions, agree on the content categories applied to the material analyzed. The more the agreement, the more the researcher can feel confident that the meaning of the content is not heavily dependent on the particular person doing the analysis. If different coders disagree frequently, then the content categories have not been defined with enough clarity and precision. Computer software greatly improves intercoder reliability because the computer performs many of the tasks that coders may otherwise need to take on themselves, reducing human error, and can be used to assist coders in following coding protocol.

Suppose we were coding the presence of Hispanics in televised entertainment programming. For each program we could count (1) whether there was at least one Hispanic present, (2) how many Hispanics there were, (3) how much time Hispanics were on the screen, and (4) how favorable the portrayal of Hispanics was or how important the portrayal of Hispanics was for the overall story. In these examples, the sampling unit and the recording unit would be the same. However, if you wanted to measure the personality traits of Hispanic prime-time television characters—such as strength, warmth, integrity, humility, and wisdom—and the

[8] For a discussion of computer-assisted text analysis, see Krippendorff, "Computer Aids," chap. 12 in *Content Analysis: An Introduction to Its Methodology;* Daniel Riffe, Stephen Lacy, and Frederick G. Fico, "Computers," chap. 9 in *Analyzing Media Messages: Using Quantitative Content Analysis in Research,* 2nd ed. (Mahwah, N.J.: Lawrence Erlbaum Associates, 2005), 208–24; and Roel Popping, "Computer Programs for the Analysis of Texts and Transcripts," in *Text Analysis for the Social Sciences: Methods for Drawing Statistical Inferences from Text and Transcripts,* ed. Carl W. Roberts (Mahwah, N.J.: Lawrence Erlbaum Associates, 1997), 209–24.

sex, age, and occupation of those characters, the program would be the sampling unit, and the individual character would be the recording unit.

As another example, if you wanted to measure the orientation of governors toward the role of government, you might use State of the State addresses. Within each address, you could code each sentence as being either positive about government or government employees, neutral, or negative. In this case, the recording unit would be the sentence, not the whole address.

Finally, a researcher has to devise a system of enumeration for the content being coded. The presence or absence of a given content category can be measured, or the measurement may be of the "frequency with which the category appears," the "amount of space allotted to the category," or the "strength or intensity with which the category is represented."[9]

An example of how political scientists use content analysis is Jonathan Paquin and Phillipe Beauregard's article, "Shedding Light on Canada's Foreign Policy Alignment."[10] The authors examined Canadian foreign policy with regard to how its responses to crises aligned with three key allies, Britain, France, and the United States. In order to examine the alignment in foreign policy between Canada and its allies, Paquin and Beauregard analyzed responses to six international crises between 2004 and 2011. According to the authors, "Alignment can be said to occur when a government publicly adopts the position of another government after the fact, or, to state it differently, when a state modifies or updates its position in order to 'bring it in line with that of another.'"[11] The authors sought to test a series of hypotheses in regard to competing theoretical explanations about Canada's foreign policy positions. First, they sought to test whether Canadian foreign policy was aligned with its allies or if Canada tended to act unilaterally without regard for its allies' positions. Second, if Canada acted in alignment with its allies, were some allies more important than others? Canada could be taking positions along continental lines, with the United States. Or Canada could align along transatlantic lines, with its European allies, Britain and France. Finally, scholars have also theorized that Canada could take positions as part of an Anglosphere with Britain and the United States, standing with its English-speaking allies.

To test their hypotheses, Paquin and Beauregard analyzed foreign policy positions from each of the four nations under study in response to the 2004 Ukrainian crises, or Orange Revolution; the 2005 assassination of former Lebanese prime

[9] Bailey, *Methods of Social Research*, 319.

[10] Jonathan Paquin and Phillipe Beauregard, "Shedding Light on Canada's Foreign Policy," *Canadian Journal of Political Science* 46, no. 3 (2013): 617–43.

[11] Ibid., 618.

minister Hariri, which led to the Cedar Revolution; the 2006 war between Israel and Hezbollah; the 2008 conflict between Russia and Georgia over South Ossetia and Abkhazia; and the 2011 Arab Spring uprisings in Egypt and Libya. Their analysis began with the first position released by a head of state or the diplomatic agency from any of the four nations under study in response to each crisis. The data included "official statements, press briefings, public letters and interviews that officially appeared on the websites and in the archives of the selected agencies."[12] In all, the authors examined 570 policy statements.

In order to test the alignment between Canada its allies, Paquin and Beauregard coded each statement during each crisis based on its date and time so they could determine which country announced its position in chronological order. Next, they coded the positions taken in each statement yielding forty-seven unique positions across the six crises. As explained earlier in the chapter, when carrying out a content analysis, researchers face the choice between manual coding and relying on computer software for assistance. In this article, the authors used a content analysis software package called QDA Miner to code the foreign policy substance of each statement. Once the positions were coded, the authors could identify instances when a nation changed its position in a way that aligned with the position of another nation. "For instance, if Canada issued a statement after the United States, France and Britain, and adopted the same position as theirs, Canada's response was then counted as aligned with each of these states."[13]

In a second analysis in the same article, Paquin and Beauregard's content analysis of foreign policy statements included coding each statement for the inclusion of forty-five different foreign policy themes. For example, some of the themes the authors coded were references to democracy, minority rights, sanctions, intervention, and self-determination. Before coding the statements the authors carefully defined each term and established rules for how they would identify the inclusion of each theme in a statement using their content analysis software. Making these definitions and rules is critical to content analysis because the definitions and rules dictate which passages would be identified for coding and which would be passed over. Across the 570 statements, the authors found 13,130 unique mentions of the forty-five themes.

This content analysis produced the data necessary to determine that Canada's foreign policy had a transatlantic orientation during the six international crises. During the period under study Canada was rarely the first of the four allies to take a foreign policy position and often aligned its foreign policy positions with Britain and France before the United States announced a position.

[12] Ibid., 626.

[13] Ibid., 626.

A Simple Computer Content Analysis

While content analysis software is likely beyond the budget for most undergraduate students, you can use your Internet browser to find and analyze speeches or other printed records if you keep careful records of your work. As an example, suppose you wanted to compare the attention leaders from different political parties gave the foreign affairs in Britain from 2005 through 2014. You could begin your analysis by reading party leader speeches archived at BritishPoliticalSpeech.org, a Web site created by scholars at Swansea University in Wales to support research on political rhetoric in Britain. The Web site offers select transcripts of leader speeches from 1895 to the present.[14] For each archived speech, you could use the browser's "Find in page" feature to look for the words *foreign, allies, Syria,* or related terms. The "Find in page" feature will highlight the word in the document and give you the number of times the word is found. For example, see figure 9-1 for an excerpt of a speech where the keyword *Syria* has been highlighted. You can record the keyword count for each term and other information about the speeches in a data matrix. We could add more variables to the data matrix—for example, the party of the leader—if we thought that this variable might influence what issues or themes leaders emphasized. A portion of such a data matrix is included in figure 9-2.

Types of Written Records

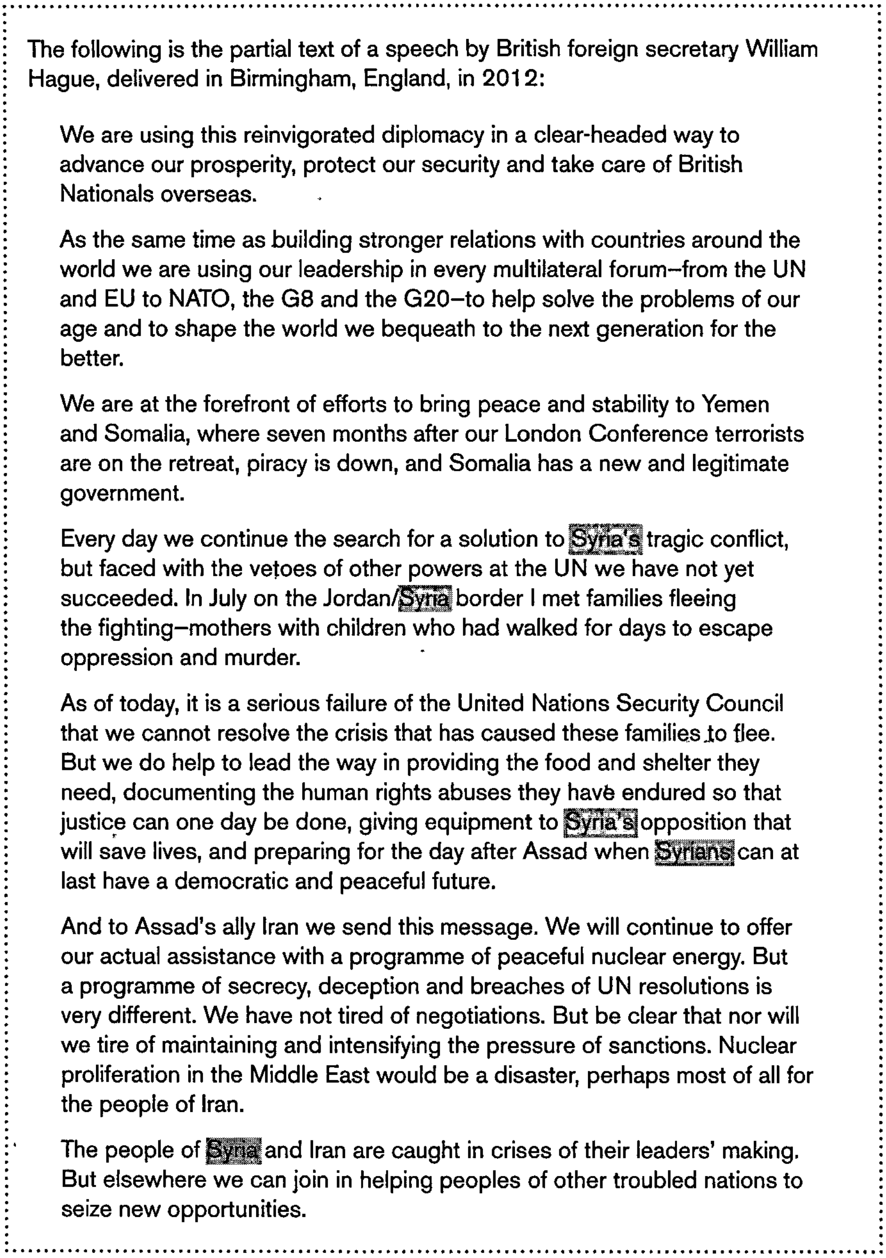
Some written records are ongoing and cover an extensive period; others are more episodic. Some are produced by public organizations at taxpayers' expense; others are produced by business concerns or by private citizens. Some are carefully preserved and indexed; others are written and forgotten. In this section, we discuss two types of written records: the running record and the episodic record.

The Running Record

The **running record** includes materials that are collected systematically across time. These records are likely to be produced by organizations rather than by private citizens, carefully stored and easily accessed, and made available for long periods of time. Governmental organizations are by far the most common source of political document collections, and these records are both extensive and growing. The increasing worldwide availability of the Internet has opened many sources of data that formerly may have been difficult or impossible to access without extensive travel.

[14] The archive is located at http://www.britishpoliticalspeech.org/speech-archive.htm

FIGURE 9-1 **Example of Search for *Syria* in 2012 William Hague Speech**



The following is the partial text of a speech by British foreign secretary William Hague, delivered in Birmingham, England, in 2012:

> We are using this reinvigorated diplomacy in a clear-headed way to advance our prosperity, protect our security and take care of British Nationals overseas.
>
> As the same time as building stronger relations with countries around the world we are using our leadership in every multilateral forum–from the UN and EU to NATO, the G8 and the G20–to help solve the problems of our age and to shape the world we bequeath to the next generation for the better.
>
> We are at the forefront of efforts to bring peace and stability to Yemen and Somalia, where seven months after our London Conference terrorists are on the retreat, piracy is down, and Somalia has a new and legitimate government.
>
> Every day we continue the search for a solution to Syria's tragic conflict, but faced with the vetoes of other powers at the UN we have not yet succeeded. In July on the Jordan/Syria border I met families fleeing the fighting–mothers with children who had walked for days to escape oppression and murder.
>
> As of today, it is a serious failure of the United Nations Security Council that we cannot resolve the crisis that has caused these families to flee. But we do help to lead the way in providing the food and shelter they need, documenting the human rights abuses they have endured so that justice can one day be done, giving equipment to Syria's opposition that will save lives, and preparing for the day after Assad when Syrians can at last have a democratic and peaceful future.
>
> And to Assad's ally Iran we send this message. We will continue to offer our actual assistance with a programme of peaceful nuclear energy. But a programme of secrecy, deception and breaches of UN resolutions is very different. We have not tired of negotiations. But be clear that nor will we tire of maintaining and intensifying the pressure of sanctions. Nuclear proliferation in the Middle East would be a disaster, perhaps most of all for the people of Iran.
>
> The people of Syria and Iran are caught in crises of their leaders' making. But elsewhere we can join in helping peoples of other troubled nations to seize new opportunities.

Source: Adapted from William Hague, Foreign Secretary's Speech in Birmingham, 2012. Available at http://www.britishpoliticalspeech.org/speech-archive.htm?speech=347

FIGURE 9-2 **Example Data Matrix Showing Results of Content Analysis of British Leader Speeches**

Case	Year	Leader	Party	Foreign	Allies/ alliances	Syria
1	2012	Hague	Conservative	10	4	6
2	2011	Cameron	Conservative	3	1	0
3	2013	Miliband	Labour	2	0	1
4	2013	Cameron	Conservative	2	0	3

Source: Compiled by the author from BritishPoliticalSpeech.org. Available at http://www.britishpoliticalspeech.org/speech-archive.htm

The running record is available for a wide variety of political topics. You can find a wealth of information related to foreign affairs from the United Nations. The UN collects and distributes data on most nations through its data Web site.[15] The UN maintains databases on major topical areas like crime, environment, labor, and many others. Alternatively, you could access data from regional data sources like the European Union. Students wishing to study the EU can quickly access data on a host of political, economic, social, and geographic data through the European Union Open Data Portal.[16] The portal is managed by the publications office of the European Union and provides access to data provided by its member nations. Another popular source of the running record in world affairs is the Central Intelligence Agency's *World Fact Book*.[17]

Domestically, the data collection and reporting efforts of the US government alone are impressive, and if you add to that the written records collected and preserved by state and local governments, interest groups, publishing houses, research institutes, and commercial concerns, the quantity of politically relevant written records increases quickly. Reports of the US government, for example, now cover everything from electoral votes to electrical rates, and taxes to taxi cabs.

If you are interested in elections and campaigns, you can visit the Federal Election Commission at www.fec.gov and find financial records filed by candidates, interest groups, and political parties, or you can visit privately operated Web sites, like

[15] Located at http://data.un.org/Default.aspx

[16] Located at https://open-data.europa.eu/en/data/

[17] Located at www.cia.gov/library/publications/the-world-factbook/index.html

www.opensecrets.org, that offer processed reports in an easy-to-read and -use format. Or you might visit the Web sites administered by the secretaries of state to find state-level election returns or summaries of election law changes over time or the America Votes series to find election results for national and some state and local elections. Alternatively, if you are interested in the lawmaking process, Congress makes the text and legislative histories of bills, committee reports, hearings, congressional votes, and the *Congressional Record* available at www.congress.gov, with a useful search engine to find needed documents. Or you can search for similar material through nongovernmental sources like the Inter-university Consortium for Political and Social Research archive or in print in the *CQ Almanac* or in CQ Press's *Politics in America.*

There is so much data available on the Internet that it may be somewhat overwhelming to try to begin a search of the running record. A good starting point may be a more general data archive like Cornell University's Institute for Social and Economic Research.[18] Students can easily identify and access data directly from this archive on a host of topics and locations across time. As you can imagine, the references listed here represent only a small fraction of the available records. Each reference has its own advantages and disadvantages, and you should take care to understand exactly what is and what is not included in each reference before using it.

THE POLICY AGENDAS PROJECT. An example of work done using the running record is John E. Uscinski's analysis of agenda setting in "When Does the Public's Issue Agenda Affect the Media's Issue Agenda (and Vice-Versa)? Developing a Framework for Media-Public Influence."[19] In this article Uscinski tested agenda-setting relationships by analyzing thirty-five thousand news stories from nightly network news broadcasts on ABC, CBS, and NBC. As we discussed above, a critical part of content analysis is defining the keywords and concepts the author is studying. For this study, the author relied on the policy topic codes developed for the Policy Agendas Project, which offers many sources of data linked by public policy topics, and is a significant resource for those doing research in public policy. [20] The Policy Agendas Project seeks to provide users with an easy, one-source way to track long-term policy changes at the national level of government across many different arenas. At the heart of this project is a comprehensive list of twenty-one major public policy topics (table 9-1). Each of the twenty-one topics is also divided into dozens of subtopics to better organize the broad policy areas.

[18] Located at http://www.ciser.cornell.edu/ASPs/datasource.asp

[19] John E. Uscinski, "When Does the Public's Issue Agenda Affect the Media's Issue Agenda (and Vice-Versa)? Developing a Framework for Media-Public Influence," *Social Science Quarterly* 90, 4 (2009): 796–815.

[20] Located at www.policyagendas.org

Finally, each topic and subtopic is assigned a unique identification number that is used in each of the datasets available on the Web site. This means that researchers can easily use data from different datasets in the archive to study public policy because each dataset uses the same coding system to identify policy topics.[21]

In this article, Uscinski used the Policy Agendas Project policy topic codes to code each news story for its policy content to establish the media agenda. He then used data available through the Policy Agendas Project to establish the public's agenda. The data he used were drawn from Gallup's "most important problem" question, a question that asks respondents to identify the most important problem facing the United States. This question has been asked by Gallup for many decades and allowed the author to follow the most important issue over time. The most important problem dataset is one of several available datasets in six distinct areas, as shown in table 9-2. Each of these datasets is a useful source of data in its own right, but the policy codes linking these data make this Web site especially important. The reason is that by using the policy codes provided by the Policy Agendas Project, the author was able to seamlessly analyze connections between his measure of media coverage and Gallup's most important question data. The same could be done with any of the data in this collection of data.

Prior research found that "reporters respond to large, spectacular, or easily reportable singular events because of their 'newsworthiness.'"[22] Using his running record data, Uscinski found that the type of event drove both media and public agenda setting. Issue areas that are commonly associated with "newsworthiness" were reported at high rates and had an agenda-setting effect on the public. Issue areas that were typically not associated with "newsworthiness" affected media coverage only when there was great public interest—the public set the media agenda in these areas.

The Episodic Record

Records that are not part of an ongoing, systematic record-keeping program but are produced and preserved in a more casual, personal, and accidental manner are called **episodic records**. Good examples are personal diaries, memoirs, manuscripts, correspondence, and autobiographies; biographical sketches and other biographical materials; the temporary records of organizations; and media of temporary existence, such as brochures, posters, and pamphlets. The episodic record

[21] In addition to the main Policy Agendas site, there are several partner sites. The Congressional Bills Project (www.congressionalbills.org), created by E. Scott Adler and John Wilkerson at the University of Washington, includes data on every congressional bill from 1947 to 2008. This Web site uses the same policy codes used on the Policy Agendas Project so that data from both sites may be easily combined. The Comparative Agendas Project (www.comparativeagendas.org) extends the Policy Agendas Project to the European Union and thirteen countries in addition to the United States. The project also includes data on two US states, Florida and Pennsylvania.

[22] Uscinski, "When Does the Public's Issue Agenda Affect the Media's Issue Agenda," 811.

TABLE 9-1 Policy Agendas Project Policy Topics

1.	Macroeconomics
2.	Civil Rights, Minority Issues, and Civil Liberties
3.	Health
4.	Agriculture
5.	Labor, Employment, and Immigration
6.	Education
7.	Environment
8.	Energy
10.	Transportation
12.	Law, Crime, and Family Issues
13.	Social Welfare
14.	Community Development and Housing Issues
15.	Banking, Finance, and Domestic Commerce
16.	Defense
18.	Space, Science, Technology, and Communications
19.	Foreign Trade
10.	International Affairs and Foreign Aid
20.	Government Operations
21.	Public Lands and Water Management
Major Topic Codes Greater than 21 (Additional NYT [*New York Times*] Codes)	
24.	State and Local Government Administration
26.	Weather and Natural Disasters
27.	Fires
28.	Arts and Entertainment
29.	Sports and Recreation
30.	Death Notices
31.	Churches and Religion
99.	Other, Miscellaneous, and Human Interest

Source: Policy Agendas Project, "Topic Codebook." Accessed August 18, 2011. Available at http://www.policyagendas.org/page/topic-codebook/

is of particular importance to political historians, since much of their subject matter can be studied only through these data.

The papers and memoirs of high-profile leaders like presidents or prime ministers could also be classified as part of the episodic record, even though considerable resources and organizational effort are invested in their preservation, insofar as the content and methods of organization of these documents vary and the papers are not all available in the same location.

To use written records, researchers must first gain access to the materials. Gaining access to the episodic record is sometimes particularly difficult.[23] Locating suitable materials can easily be the most time-consuming aspect of the whole data collection exercise.

Researchers generally use episodic records to illustrate phenomena rather than as a basis for the generation of a large sample and numerical measures for statistical analysis. Consequently, quotations and other excerpts from research materials are often used as evidence for a thesis or hypothesis. That is to say, their analyses are qualitative rather than quantitative. Over the years, social scientists have conducted some exceptionally interesting and imaginative studies of political phenomena based on the episodic record.

PRESIDENTIAL PERSONALITY. An example of the use of the episodic record may be found in James David Barber's *The Presidential Character.*[24] Because of the importance of the presidency in the American political system and the extent to which that institution is shaped by its sole occupant, Barber was interested in understanding the

[23] Charles A. Beard reported that he was able to use some records in the US Treasury Department in Washington, D.C., "only after a vacuum cleaner had been brought in to excavate the ruins." See Beard, *An Economic Interpretation of the Constitution of the United States* (New York: Macmillan, 1913), 22.

[24] James David Barber, *The Presidential Character: Predicting Performance in the White House*, 3rd. ed. (Englewood Cliffs, N.J.: Prentice Hall, 1985).

TABLE 9-2 Policy Agendas Project Datasets

Congress

Congressional Hearings

This dataset contains information summarizing each U.S. Congressional hearing from 1946 to 2010 (91,656 hearings). Using the Congressional Information Service (CIS) Abstracts, we code each hearing by our system of policy content codes. Other variables, including committee and subcommittee, are also available. Identification variables link our records to the original CIS source material. Note: Research making use of the congressional hearings dataset should bear in mind that the hearings for the last year available on our website are incomplete. This is due to the CIS archival system.

Congressional Quarterly Almanac

This dataset contains information from all articles in the main chapters of the CQ Almanac from 1948 to 2011 (14,217 records). Each CQ Almanac articles typically covers one legislative initiative; when an article contains information about several different public laws or bills, it is divided so that each record in our dataset contains information about one legislative initiative. Each record is coded according to our policy content scheme. Several other variables concerning each legislative initiative (e.g., bill numbers, Public Law number if applicable, committees involved, primary sponsors, etc.) are also included. Identification variables link our records to the original CQ source material as well as to our Public Laws dataset. A note of caution, article length has varied over the span of this dataset.

Public Laws

This dataset contains information about each public law passed from 1948 to 2011 (19,914 records). Each record is coded by our policy content scheme and other variables. Identification variables allow linkage to the CQ Almanac dataset. The dataset directly links users to the full text (starting with the 104th Congress) and bill summary (starting with the 93rd Congress) information found on THOMAS and other public domain websites.

Roll Call Votes

The Congressional Roll Call Voting dataset codes every congressional roll call vote from 1947 to 2012 (49,216 votes) using the Policy Agendas Project content coding system. In addition, this dataset standardizes information from multiple sources into an easily utilized format. As of August 2014, we have streamlined the variables that we collect and offer for download in the RC dataset. A link to the legacy version and corresponding data codebook is available below.

Presidency

Executive Orders

This dataset contains information about each executive order issued from 1945 to 2013 (4,129 records). Each record is coded according to our policy content scheme and other variables including the presidents party, whether the order was issued during a time of divided government, and whether the order was issued at the beginning or end of a presidential term.

State of the Union Speeches

This dataset contains information on each quasi-statement in the Presidential State of the Union Speeches from 1946 to 2015 (22,417 records). Each quasi-statement is coded according to our system of policy content categories and other variables. Users can directly link to full text versions of the speech for further analysis.

Supreme Court

Supreme Court Cases

The Supreme Court dataset contains information on each case on the Courts docket from 1945 to 2009 (8,955 records), and is the only publicly available dataset to examine the Courts agenda from a policy perspective. Cases are coded according to policy content and include additional variables such as the Courts ruling in cases in which one was issued. The accompanying codebook addresses Court-specific coding issues and serves as a reference guide for those unfamiliar with the Courts terminology and procedures.

Public Opinion and Interest Groups

Encyclopedia of Associations

Since 1956, Gale Research, later Thomson/Gale, has published a printed volume entitled the Encyclopedia of Associations. The database on which the book is based also serves as a web-based research tool available through libraries and entitled Associations Unlimited. While not originally designed with the idea of dynamic analysis in mind, the accumulated volumes of the EA in fact allow a researcher considerable opportunity for analyzing trends over time. The Policy Agendas Project (PAP) has used the annual volumes of the EA to compile a time-series

(Continued)

TABLE 9-2 (Continued)

database of all associations, coded both by the EA subject categories as well as by the major topics of the PAP. Forty-two editions of the EA have been published from 1956 to 2005. We have compiled a simple list of each group and coded it into the PAP topic classification system. Complete data are available in 5-year intervals from 1970-2005 as well as estimated annual counts for the full period. A description of coverage and important details concerning the lag between reported copyright years and the information they represent is included in the full dataset codebook. Note that as of March 2014, we have implemented a 4 year lag in the annual dataset, with the previous Year variable now listed as CopyrightYear. Below are links to the annual imputed counts dataset (1966-2001, 972 records) used in the trends analysis tool (with corresponding codebook) as well the full 1970-2005 dataset (with corresponding codebook). A recently published article about the dataset is also provided.

Gallup's Most Important Problem

This dataset contains responses to Gallup's Most Important Problem question aggregated at the annual level from 1946 to 2012 (1,407 records) and coded by major topic. Years with missing observations (1953/1955) are those in which there were no corresponding MIP data available. Contact us for quarterly MIP data if needed.

Policy Moods

The policy specific moods data set, compiled by James A. Stimson and K. Elizabeth Coggins, was created to supplement the traditional Global Mood measure in an effort to provide scholars with as many policy specific mood measures as possible. The global mood database, which consists of nearly 400 survey questions and almost 8,000 administrations across 70 years, was disaggregated to generate longitudinal measures of public opinion in specific policy domains. By matching each survey item with a policy code from the Policy Agendas Project coding scheme, it was possible to estimate 61 unique series as well as five additional series relating to abortion and gay rights spanning 1946 to 2011 (3,099 records). More information about survey items, administrations and time periods can be found in the corresponding data codebook.

News Media

New York Times Index

This dataset is a systematic random sample of the New York Times Index from 1946 to 2008 (49,201 records). The sample includes the first entry on every odd-numbered page of the Index. Each entry is coded by Policy Agendas major topics and includes other variables such as the length, date and location of the story and whether it addressed government actions.

New York Times Index Weights

This dataset provides information on the number of pages in the New York Times Index and an estimate of the number of articles per page for each of the years included in our Index dataset. These weights address the occasional newspaper format changes that systematically alter the number of articles on each page and the variation in the size of the New York Times and its Index over time.

Federal Budget

Budget Authority Adjusted

This dataset provides annual data, adjusted for inflation, of U.S. Budget Authority from FY 1947 through FY 2014 (7,820 records). Using Office of Management and Budget Functions and Subfunctions, we have revised the data to be consistent across time. We utilize the most recent OMB deflator to generate inflation-adjusted variables.

Budget Authority-Policy Crosswalk

This file compares the Policy Agendas Project topic codes with the OMB codes used in the Budget Authority dataset to assess how well they correspond. A "1" represents nearly complete correspondence, while a "5" represents significant divergence.

Budget Outlays

This dataset, compiled by Bryan D. Jones, Frank R. Baumgartner and John Lovett, provides two 'synthetic' series of annual, long-term budget outlays. There is no single series reporting expenditures (outlays) for the US Federal Government since the founding of the Republic. However, two separate data series are available for US Federal Expenditures, compiled by the Treasury Department and Office of Management and Budget. The Treasury Series runs from 1791 to 1970, and the OMB series covers 1940 to the present. From these data sources, two synthetic budget series are constructed by merging data from the US Treasury with data from OMB. The series labeled Treasury Synthetic uses Treasury data from 1791 through 1970, OMB afterward. OMB Synthetic uses Treasury numbers until 1940, OMB afterward. For a complete description of these data sources, methods used to construct the series, and variable descriptions, please see the corresponding codebook below.

Budget Resources

These pages highlight the main issues concerning the study of budgetary outcomes across countries and time. A brief glossary of budgetary terminology and data sources from international, national, and research institutions are provided.

Source: Policy Agendas Project, "Datasets & Codebooks." Accessed August 14, 2015. Available at http://www.policyagendas.org/page/datasets-codebooks#codebook personalities of the individuals who had occupied the office during the twentieth century. Although he undoubtedly would have preferred to observe directly the behavior of the fourteen presidents who held office between 1908 and 1984 (when he conducted his study), he was forced instead to rely on the available written materials about them.

For Barber, discerning a president's personality meant understanding his style, worldview, and character. Style is "the President's habitual way of performing his three political roles: rhetoric, personal relations, and homework." A president's worldview is measured by his "primary, politically relevant beliefs, particularly his conceptions of social causality, human nature, and the central moral conflicts of the time." And character "is the way the President orients himself toward life." Barber believed that a president's style, character, and worldview "fit together in a dynamic package understandable in psychological terms" and that this personality "is an important shaper of his Presidential behavior on nontrivial matters." But how is one to measure the style, character, and worldview of presidents who are dead or who will not permit a political psychologist access to their thoughts and deeds? This is an especially troublesome question when one believes, as Barber did, that "the best way to predict a President's character, world view, and style is to see how they were put together in the first place . . . in his early life, culminating in his first independent political success."[25]

Barber's solution to this problem was to use available materials on the twentieth-century presidents he studied, including biographies, memoirs, diaries, speeches, and, for Richard Nixon, tape recordings of presidential conversations. Barber did not use all the available biographical materials. For example, he "steered clear of obvious puff jobs put out in campaigns and of the quickie exposés composed to destroy reputations."[26] He quoted frequently from the biographical materials as he built his case that a particular president was one of four basic personality types. Had these materials been unavailable or of questionable accuracy (a possibility that Barber glosses over in a single paragraph), measuring presidential personalities would have been a good deal more difficult, if not impossible.

Barber's analysis of the presidential personality was exclusively qualitative; the book contains not one table or graph. He used the biographical material to categorize each president as one of four personality types and to show that the presidents with similar personalities exhibited similar behavioral patterns when in office. In brief, Barber used two dimensions—activity-passivity (how much energy does the man invest in his presidency?) and positive-negative affect (how does he feel about what he does?)—to define the four types of presidential personality (table 9-3).

[25] Ibid., 4–5

[26] James David Barber, *The Presidential Character: Predicting Performance in the White House* (Englewood Cliffs, N.J.: Prentice-Hall, 1972), ix.

TABLE 9-3 **Presidential Personality Types**

Positive-negative affect	Activity-Passivity	
	Active	Passive
Positive	Franklin D. Roosevelt Harry S. Truman John F. Kennedy Gerald Ford Jimmy Carter	William Howard Taft Warren Harding Ronald Reagan
Negative	Woodrow Wilson Herbert Hoover Lyndon Johnson Richard Nixon	Calvin Coolidge Dwight Eisenhower

Source: Based on data from James David Barber, *The Presidential Character*, 3rd ed. (Englewood Cliffs, N.J.: Prentice Hall, 1985). Courtesy of James David Barber, James B. Duke Professor of Science, Emeritus, Duke University, Durham, N.C.

Barber's research is a provocative and imaginative example of the use of the episodic record—in this case, biographical material—as evidence for a series of generalizations about presidential personality. Although Barber did not empirically test his hypotheses in the ways that we have been discussing in this book, he did accumulate a body of evidence in support of his assertions and presented his evidence in such a way that the reader can evaluate how persuasive it is.[27]

The Running Record and Episodic Record Compared

There are three primary advantages to using the running record rather than the episodic record. The first is cost, in both time and money. Since the costs of collecting, tabulating, storing, and reporting the data in the running record are generally borne by the record keepers themselves, political scientists are usually able to use these data inexpensively. Researchers can often use the data stored in the running record by photocopying a few pages of a reference book, purchasing a government report or data file, or downloading data into a spreadsheet. In fact, the continued expansion of the data collection and record-keeping activities of national governments has been a financial boon to social scientists of all types.

A second, related advantage is the accessibility of the running record. Instead of searching packing crates, deteriorated ledgers, and musty storerooms, as users of the episodic record often must do, users of the running record more often rely on downloading data files and handling reference books and government publications. Many political science research projects have been completed with only the data stored in the reference books and government documents of a decent research library or through online archives.

A third advantage of the running record is that, by definition, it covers a more extensive period than does the episodic record. This permits the type of longitudinal analysis and before-and-after research designs discussed in chapter 6 and in the agenda-setting example above. Although the episodic record helps explain the

[27] A critique of Barber's analysis may be found in Garry Wills, *The Kennedy Imprisonment: A Meditation on Power* (Boston: Little, Brown, 1982).

origins of and reasons for a particular event, episode, or period, the running record allows the measurement of political phenomena over time.

The running record presents problems, however. One is that a researcher is at the mercy of the data collection practices and procedures of the record-keeping organizations themselves. Researchers are rarely in a position to influence record-keeping practices; they must rely instead on what organizations such as the US Census Bureau, the European Union, and the Policy Agendas Project decide to do. A trade-off often exists between ease of access and researcher influence over the measurements that are made. Some organizations—some state and local governments, for example—do not maintain records as consistently as researchers may like. One colleague found tracing the fate of proposed constitutional amendments to the Delaware State Constitution to be a difficult task. Delaware is the only state in which voters do not ratify constitutional amendments. Instead, the state legislature must pass an amendment in two consecutive legislative sessions between which a legislative election has occurred. Thus, constitutional amendments are treated like bills, and tracking them depends on the archival practices of the state legislature. Even when clear records are kept, such as election returns for mayoral contests, researchers may face a substantial task in collecting the data from individual cities, because the only returns from the largest cities are reported in various statistical compilations.

Another, related disadvantage of the running record is that some organizations are not willing to share their raw data with researchers. The processed data that they do release may reflect calculations, categorizations, and aggregations that are inaccurate or uninformative. Access to public information is not *always* easy. More problems may be encountered when trying to obtain public information that shares some of the characteristics of the episodic record, for example, such as information on the effect of specific public programs and agency activities. Emily Van Dunk, a senior researcher at the Public Policy Forum, a nonpartisan, nonprofit research organization that conducts research on issues of importance to Wisconsin residents, noted that obtaining data from state and local government agencies can be difficult at times and offered tips for researchers.[28]

Finally, it is sometimes difficult for researchers to find out exactly what an organization's record-keeping practices are. Unless the organization publishes a description of its procedures, a researcher may not know what decisions have guided the record-keeping process. This can be a special problem when these practices change, altering in an unknown way the measurements reported.

Although the running record has its disadvantages, political scientists often must rely on it if they wish to do any empirical research on a particular topic. To illustrate

[28] Emily Van Dunk, "Getting Data through the Back Door: Techniques for Gathering Data from State Agencies," *State Politics and Policy Quarterly* 1, no. 2 (2001): 210–18.

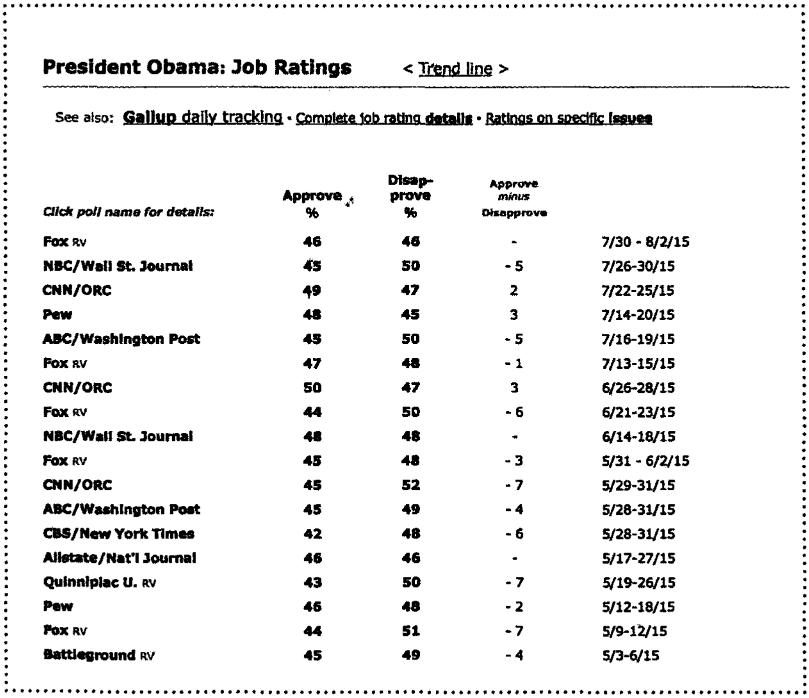
some of the problems with using written records, we conclude this section with a description of PollingReport.com, one of many Web sites dedicated to providing users with national and state-level public opinion data.

Presidential Job Approval

PollingReport.com is a popular source of public opinion polling data. PollingReport.com provides national poll results, free of charge, from well-known polling organizations such as Gallup, Pew, and Quinnipiac and news organizations such as CNN, CBS, and the *Los Angeles Times*. The Web site also offers state-level poll results to paid subscribers. In this section, we focus on the data available for free.

PollingReport.com organizes its poll results into the following categories: the State of the Union, Elections, In the News, National Security, and Issues. Each of these categories offers a number of subtopics of interest. The State of the Union category, for example, includes subtopics covering each branch of the federal government—"President Obama," "Congress," and "Supreme Court"—as well as "Direction of the country" and "National priorities." A great deal of useful public opinion data may be found among these many subtopics.

FIGURE 9-3 Presidential Support Data at PollingReport.com



President Obama: Job Ratings < Trend line >

See also: Gallup daily tracking · Complete job rating details · Ratings on specific issues

Click poll name for details:	Approve %	Disap-prove %	Approve *minus* Disapprove	
Fox RV	46	46	-	7/30 - 8/2/15
NBC/Wall St. Journal	45	50	- 5	7/26-30/15
CNN/ORC	49	47	2	7/22-25/15
Pew	48	45	3	7/14-20/15
ABC/Washington Post	45	50	- 5	7/16-19/15
Fox RV	47	48	- 1	7/13-15/15
CNN/ORC	50	47	3	6/26-28/15
Fox RV	44	50	- 6	6/21-23/15
NBC/Wall St. Journal	48	48	-	6/14-18/15
Fox RV	45	48	- 3	5/31 - 6/2/15
CNN/ORC	45	52	- 7	5/29-31/15
ABC/Washington Post	45	49	- 4	5/28-31/15
CBS/New York Times	42	48	- 6	5/28-31/15
Allstate/Nat'l Journal	46	46	-	5/17-27/15
Quinnipiac U. RV	43	50	- 7	5/19-26/15
Pew	46	48	- 2	5/12-18/15
Pox RV	44	51	- 7	5/9-12/15
Battleground RV	45	49	- 4	5/3-6/15

Source: PollingReport.com, "President Obama: Job Ratings." Available at http://www.pollingreport.com/obama.htm

Let's assume that you are interested in studying public support for the president. The place to start would be finding data on President Obama's job approval ratings, which are perhaps the most direct indicator of support for the president. By clicking on the President Obama subtopic under State of the Union, you will find several different kinds of presidential support data (figure 9-3). PollingReport.com provides data on the "Obama administration," "Job ratings," and "Favorability ratings." You will want to explore each of these options, looking for differences in polling questions and responses. We will explore President Obama's job ratings poll results in this example. You can click on

"Major polls" to find poll results from various polling organizations that answer a question similar to the ABC News/*Washington Post* poll's job approval rating question: "Do you approve or disapprove of the way Barack Obama is handling his job as president?" or "Daily tracking," which will give you the results of the Gallup Poll question asked virtually daily since January 2009. A portion of these results is shown in figure 9-4. From this page, you can access polling results for questions asking about President Obama's handling of specific issues. These data could be used to investigate how opinions about the handling of specific issues affect assessments of overall job performance.

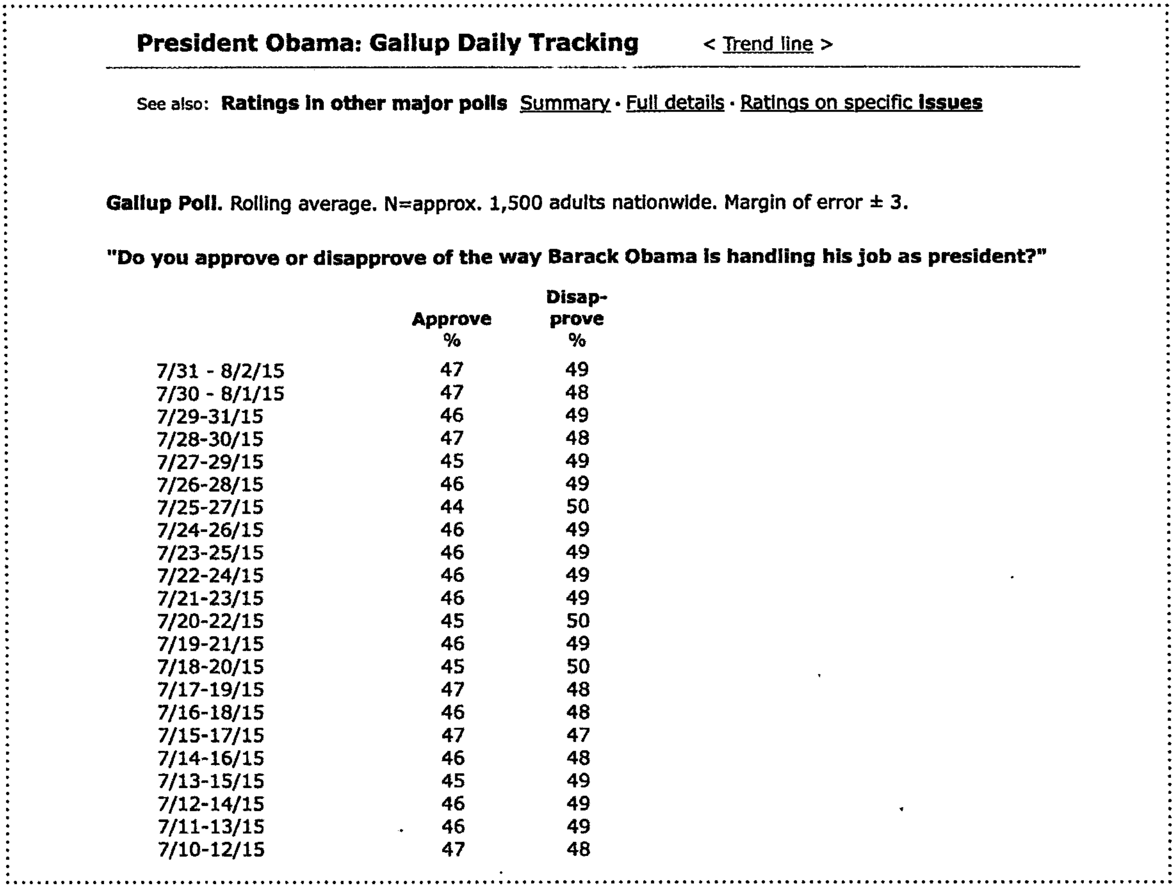
PollingReport.com has many advantages for students using the written record. First, and perhaps most important, PollingReport.com offers free, high-quality data at an easy-to-use Web site. The results found at PollingReport.com come from the same professional polling organizations on which news organizations around the country rely. Second, students have access to multiple surveys administered during different periods using very similar question wording. But as valuable as PollingReport.com is, it shares some disadvantages with other examples of the running record. Perhaps most glaring is the lack of consistency and regularity in the poll results provided on the Web page. Even though the president's job approval rating question is one of the most frequently asked questions in national political surveys, other questions are not asked with similar frequency or duration. This is not an indictment of PollingReport.com but a symptom of the fact that PollingReport.com can report only the data made available by other polling organizations. Although those other organizations provide a great deal of data, sometimes a large number of surveys are administered at the same time whereas no data are available for other time periods. And, although a great many organizations are listed on the Web site, it does not include all polling organizations.

Finally, PollingReport.com provides poll results from the Bush and Clinton administrations, but results from previous administrations are not available. If you wish to compare President Obama's approval ratings with those of other previous presidents, you will have to search for those results elsewhere. These are only some of the potential problems that might be encountered with PollingReport.com or other examples of the running record. Problems like these generally will not prevent you from using such sources, but they can be a nuisance depending on the purpose of your research project.

Advantages and Disadvantages of the Written Record

Using documents and records, or what we have called the written record, has several advantages for researchers. We highlight six of the advantages here. First, it allows us access to subjects that may be difficult or impossible to research through

FIGURE 9-4 Gallup Daily Tracking Data for President Obama from PollingReport.com



President Obama: Gallup Daily Tracking < Trend line >

See also: **Ratings in other major polls** Summary · Full details · Ratings on specific **issues**

Gallup Poll. Rolling average. N=approx. 1,500 adults nationwide. Margin of error ± 3.

"Do you approve or disapprove of the way Barack Obama is handling his job as president?"

	Approve **%**	**Disap-prove** **%**
7/31 - 8/2/15	47	49
7/30 - 8/1/15	47	48
7/29-31/15	46	49
7/28-30/15	47	48
7/27-29/15	45	49
7/26-28/15	46	49
7/25-27/15	44	50
7/24-26/15	46	49
7/23-25/15	46	49
7/22-24/15	46	49
7/21-23/15	46	49
7/20-22/15	45	50
7/19-21/15	46	49
7/18-20/15	45	50
7/17-19/15	47	48
7/16-18/15	46	48
7/15-17/15	47	47
7/14-16/15	46	48
7/13-15/15	45	49
7/12-14/15	46	49
7/11-13/15	46	49
7/10-12/15	47	48

Source: PollingReport.com, "President Obama: Job Ratings." Available at http://www.pollingreport.com/obama_job.htm

direct, personal contact because they pertain either to the past or to phenomena that are geographically distant. For example, late-eighteenth-century records permitted Charles Beard to advance and test a novel interpretation of the framing of the US Constitution. Beard suggested that the framers wrote the Constitution with their own economic interests in mind and found evidence for his argument in disparate records from the period including biographical materials, US Treasury and Census records, state loan officer and business records, and personal papers stored in the Library of Congress.[29] This study would not have been possible had no records been available from this period.

[29] Beard, *An Economic Interpretation of the Constitution of the United States.*

A second advantage of data gleaned from archival sources is that the raw data are usually nonreactive. As we mentioned in previous chapters, human subjects often consciously or unconsciously establish expectations or other relationships with investigators, which can influence their behavior in ways that might confound the results of a study. But those writing and preserving the records are frequently unaware of any future research goal or hypothesis or, for that matter, that the fruits of their labors will be used for research purposes at all. State loan officers during the late 1700s had no idea that some two hundred years later, a historian would use their records to discover why some people were in favor of revising the Articles of Confederation. This nonreactivity has the virtue of encouraging more accurate and less self-serving measures of political phenomena.

Record keeping is not always completely nonreactive, however. Record keepers are less likely to create and preserve records that are embarrassing to them, their friends, or their bosses; that reveal illegal or immoral actions; or that disclose stupidity, greed, or other unappealing attributes. Richard Nixon, for example, undoubtedly wished that he had destroyed or never made the infamous Watergate tapes, which revealed the extent of his administration's knowledge of the 1972 break-in at Democratic National Committee headquarters. Today many record-keeping agencies employ paper shredders to ensure that a portion of the written record does *not* endure. Researchers must be aware of the possibility that the written record has been selectively preserved to serve the record keepers' own interests.

A third advantage of using the written record is that sometimes the record has existed long enough to permit analyses of political phenomena over time. The before-and-after research designs discussed in chapter 6 may then be used. For example, suppose you are interested in how changes in the fifty-five-mile-per-hour speed limit (gradually adopted by the states and then later dropped by many states on large stretches of their highway systems) affected the rate of traffic accidents. Assuming that the written record contains data on the incidence of traffic accidents over time in each state, you could compare the accident rate before and after changes in the speed limit in those states that changed their speed limit. These changes in the accident rate could then be compared with the changes occurring in states in which no change in the speed limit took place. The rate changes could then be "corrected" for other factors that might affect the rate of traffic accidents. In this way an interrupted time series research design could be used, a research design that has some important advantages over cross-sectional designs. Because of the importance of time, and of changes in phenomena over time, for the acquisition of causal knowledge, a data source that supports longitudinal analyses is a valuable one. The written record more readily permits longitudinal analyses than do either interview data or direct observation.

A fourth advantage to researchers of using the written record is that it often enables us to increase sample size above what would be possible through either interviews or direct observation. For example, it would be terribly expensive and time-consuming to observe the level of spending by all candidates for the House of Representatives in any given year. Interviewing candidates would require a lot of travel, long-distance phone calls, or the design of a questionnaire to secure the necessary information. Direct observation would require gaining access to many campaigns. How much easier and less expensive it is to contact the Federal Election Commission in Washington, D.C., and request the printout of campaign spending for all House candidates. Without this written record, resources might permit only the inclusion of a handful of campaigns in a study; with the written record, all 435 campaigns can easily be included.

This raises the fifth main advantage of using the written record: cost. Since the cost of creating, organizing, and preserving the written record is borne by the record keepers, researchers are able to conduct research projects on a much smaller budget than would be the case if they had to bear the cost themselves. In fact, one of the major beneficiaries of the record-keeping activities of the federal government and of news organizations is the research community. It would cost a prohibitive amount for a researcher to measure the amount of crime in all cities larger than 25,000 or to collect the voting returns in all 435 congressional districts. Both pieces of information are available at little or no cost, however, because of the record-keeping activities of the FBI and the Elections Research Center, respectively. Similarly, using the written record often saves a researcher considerable time. It is usually much quicker to consult printed government documents, reference materials, computerized data, and research institute reports than it is to accumulate data ourselves. The written record is a veritable treasure trove for researchers.

A final advantage of using the written record is that it raises fewer ethical issues than either firsthand observation or interviewing. Research involving the collection or study of existing data, documents, or records often does not pose risks to individuals, because the unit of analysis for the data is not the individual. Also, issues of risk are not likely to arise where records are for individuals, as long as individuals cannot be identified directly or through identifiers linked to them (organizations often go to great lengths to delete possible personal identifying information) or where the records are publicly available, as in the case of the papers of public figures such as presidents and members of Congress. However, allowing researchers access to their private papers may pose some risk to private individuals. Thus, access to private papers may be subject to conditions designed to protect the individuals involved.

Collecting data in this manner, however, is not without some disadvantages. We discuss five disadvantages here. One problem mentioned earlier is selective survival. For a variety of reasons, record keepers may not preserve all pertinent materials but rather selectively save those that are the least embarrassing, controversial, or problematic. It would be surprising, for example, if political candidates, campaign consultants, and public officials saved correspondence and memoranda that cast disfavor on themselves. Obviously, whenever a person is selectively preserving portions of the written record, the accuracy of what remains is suspect. This is less of a problem when the connection between the record keeper's self-interest and the subject being examined by the researcher is minimal.

A second, related disadvantage of the written record is its incompleteness. Large gaps exist in many archives due to fires, losses of other types, personnel shortages that hinder record-keeping activities, and the failure of the record maker or record keeper to regard a record as worthy of preservation. We all throw out personal records every day; political entities do the same. It is difficult to know what kinds of records should be preserved, and it is often impossible for record keepers to bear the costs of maintaining and storing voluminous amounts of material.

Another reason why records may be incomplete is simply because no person or organization has assumed the responsibility for collecting or preserving them. For example, before 1930, national crime statistics were not collected by the FBI, and before the creation of the Federal Election Commission in 1971, records on campaign expenditures by candidates for the US Congress were spotty and inaccurate.

A third disadvantage of the written record is that its content may be biased. Not only may the record be incomplete or selectively preserved, but it also may be inaccurate or falsified, either inadvertently or on purpose. Memoranda or copies of letters that were never sent may be filed, events may be conveniently forgotten or misrepresented, the authorship of documents may be disguised, and the dates of written records may be altered; furthermore, the content of government reports may tell more about political interests than empirical facts. For example, Soviet and Eastern European governments apparently released exaggerated reports of their economic performance for many years, and scholars (and investigators) attempting to reconstruct the actions in the Watergate episode have been hampered by alterations of the record by those worried about the legality of their role in it. Often, historical interpretations rest upon who said or did what, and when. To the extent that falsifications of the written record lead to erroneous conclusions, the problem of record-keeping accuracy can bias the results of a research project. The main safeguard against bias is the one used by responsible journalists: confirming important pieces of information through several dissimilar sources.

A fourth disadvantage is that some written records are unavailable to researchers. Documents may be classified by the federal government, they may be sealed (that is, not made public) until a legal action has ceased or the political actors involved have passed away, or they may be stored in such a way that they are difficult to use. Other written records—such as the memoranda of multinational corporations, campaign consultants, and Supreme Court justices—are seldom made public because there is no legal obligation to do so and the authors benefit from keeping them private.

Finally, the written record may lack a standard format because it is kept by different people. For example, the Chicago budget office may have budget categories for public expenditures different from those used in the San Francisco budget office. Or budget categories used in the Chicago budget office before 1960 may be different from the ones used after 1960. Or the French may include items in their published military defense expenditures that differ from those included by the Chileans in their published reports. Consequently, a researcher often must expend considerable effort to ensure that the formats in which the records of different entities can be made comparable.

Despite these limitations, political scientists have generally found that the advantages of using the written record outweigh the disadvantages. The written record often supplements the data we collect through interviews and direct observation, and in many cases it is the only source of data on historical and cross-cultural political phenomena.

Conclusion

The written record includes personal records, archival collections, organizational statistics, and the products of the news media. Researchers interested in historical research, or in a particular event or time in the life of a polity, generally use the episodic record. Gaining access to the appropriate material is often the most resource-consuming aspect of this method of data collection, and the hypothesis testing that results is usually more qualitative and less rigorous (some would say more flexible) than with the running record. Increasingly, gaining access is less of a problem as more and more documents are scanned and made available online.

The running record of organizations has become a rich source of political data as a result of the record-keeping activities of governments at all levels and of interest groups and research institutes concerned with public affairs. The running record is generally more quantitative than the episodic record and may be used to conduct longitudinal research. Measurements using the running record can often be obtained inexpensively, although the researcher frequently relinquishes considerable control over the data collection enterprise in exchange for this economy.

One of the ways in which a voluminous, nonnumerical written record may be turned into numerical measures and then used to test hypotheses is through a procedure called content analysis. Content analysis is most frequently used by political scientists interested in studying media content, but it has been used to advantage in studies of political speeches, statutes, and judicial decisions.

Through the written record, researchers may observe political phenomena that are geographically, physically, and temporally distant from them. Without such records, our ability to record and measure historical phenomena, cross-cultural phenomena, and political behavior that do not occur in public would be seriously hampered.

TERMS INTRODUCED

Content analysis. A systematic procedure by which records are transformed into quantitative data.

Episodic records. Record that is not part of a regular, ongoing record-keeping enterprise but instead is produced and preserved in a more casual, personal, or accidental manner.

Intercoder reliability. Demonstration that multiple analysts, following the same content analysis procedure, agree and obtain the same measurements.

Running record. A written record that is enduring and easily accessed and covers an extensive period of time.

Written record. Documents, reports, statistics, manuscripts, and other recorded materials available and useful for empirical research.

SUGGESTED READINGS

Grbich, Carol. *Qualitative Data Analysis: An Introduction.* 2nd ed. Thousand Oaks, Calif.: Sage, 2012.

Krippendorff, Klaus. *Content Analysis: An Introduction to Its Methodology.* 3rd ed. Thousand Oaks, Calif.: Sage, 2013.

Miller, Delbert C., and Neil J. Salkind. *Handbook of Research Design and Social Measurement.* 6th ed. Newbury Park, Calif.: Sage, 2002.

Riffe, Daniel, Stephen Lacy, and Frederick G. Fico. *Analyzing Media Messages: Using Quantitative Content Analysis in Research.* 2nd ed. Mahwah, N.J.: Lawrence Erlbaum Associates, 2005.

Roberts, Carl W., ed. *Text Analysis for the Social Sciences: Methods for Drawing Statistical Inferences from Text and Transcripts.* Mahwah, N.J.: Lawrence Erlbaum Associates, 1997.

Silver, Christina, and Ann Lewins. *Using Software in Qualitative Research.* 2nd ed. Thousand Oaks, Calif.: Sage, 2014.

Van Dunk, Emily. "Getting Data through the Back Door: Techniques for Gathering Data from State Agencies." *State Politics and Policy Quarterly* 1, no. 2 (2001): 210–18.

Webb, Eugene J., Donald T. Campbell, Richard J. Schwartz, and Lee Sechrest. *Unobtrusive Measures.* Rev. ed. Thousand Oaks, Calif.: Sage, 2000.

CHAPTER 10

Survey Research and Interviewing

CHAPTER OBJECTIVES

10.1 Describe the ways in which survey research and interviewing ensure validity and reliability.

10.2 Discuss the elements of survey research and their importance.

10.3 Identify the costs and benefits of using archived surveys versus conducting your own survey.

10.4 Explain the role of interviewing in survey research.

POSSIBLY THE MOST HOTLY CONTESTED ISSUE in the first decade of twenty-first-century American politics was passage of the Patient Protection and Affordable Care Act (PPACA), frequently labeled by its critics as "Obamacare" after President Barack Obama. Among many, many objections (and possibly because of them) opponents claim the law, which tries to expand health insurance coverage in the United States, is bitterly opposed by a vast majority of citizens. Typical of comments directed at the legislation is this *Washington Examiner* editorial: "Six months ago, President Obama, Senate Majority Leader Harry Reid and House Speaker Nancy Pelosi *rammed Obamacare down the throats of an unwilling American public* [emphasis added]."[1] More recently, House Speaker John Boehner[2] and new Senate majority

[1] "*Examiner* Editorial: Obamacare Is Even Worse Than Critics Thought," *Washington Examiner*, September 22, 2010. Available at http://www.washingtonexaminer.com/examiner-editorial-obamacare-is-even-worse-than-critics-thought/article/89183

[2] John Boehner and Mitch McConnell, "Now We Can Get Congress Going," *Wall Street Journal*, November 5, 2014. Available at http://www.wsj.com/articles/john-boehner-and-mitch-mcconnell-now-we-can-get-congress-going-1415232759

leader Mitch McConnell decried the law, writing that it is a "hopelessly flawed law that *Americans have never supported* [emphasis added]." Yet one can reasonably ask, how do we know that the people had to have the law "rammed" down their throats? Is it true that Americans have never supported the law? One answer is that "polls say so." That response, of course, simply raises new issues: Which polls? Who sponsored them? What do they really show? And, by the way, how reliable is polling for measuring public opinion?

Most readers are no doubt familiar with this form of argumentation because even a cursory glance at the news demonstrates how pervasive polling has become in American politics. Polls are part and parcel of the efforts of many groups not just to study public opinion but also to use it for political ends. So it behooves anyone who wants to understand debates about public policies and issues to become familiar with this activity.

More generally, polling—or, as we call it, survey research—is an indispensable tool in social and political research. Suppose we want to know whether or not Americans are "isolationists" or "internationalists" when it comes to foreign affairs. We might try to answer the question by making indirect or unobtrusive observations, such as reading letters to the editor in a dozen or so newspapers and coding them as pro or con involvement. Or we might observe protest demonstrations for or against various international activities to see what kinds of people seem to be participating. But these indirect methods probably would not tell us what we wanted to know. It would seem far preferable (and maybe even easier) to ask citizens up front how they felt about world affairs and the proper US role in them. In this chapter, we explain two related methods of collecting data from people: (1) survey research, which involves collecting information via a questionnaire or **survey instrument** (a carefully structured or scripted set of questions that may be administered face to face, by telephone, by mail, by Internet, or by other means), and (2) interviewing, which involves direct and personal communication with individuals in a less formal and less structured situation—more in the nature of a constrained conversation. Although we describe both techniques in a moment, for now let us just say that these approaches range from talking to one or a handful of people to gathering data from 1,000 or more people across an entire nation. In either approach, the researcher is trying to get at what people think and do by asking them for self-reports.

Because both methods rely on interpersonal communication, they might seem to entail no special considerations: just think of some questions and ask them. This is not the casė, however. To see why, refer to the research described in chapter 1 regarding voter turnout. As you may recall, the issue boils down to who votes and who doesn't. Political scientists have developed all sorts of hypotheses and theories to answer the question, but testing them rests on a seemingly simple and straightforward but in reality quite difficult task: determining who actually voted in any given election. You might think it would be easy to ask people, "Did you vote in the last election?" And that is precisely what most survey researchers do. The problem is that often two-thirds to three-fourths of the respondents claim to have voted. But we know from vote counts reported by election officials that these survey estimates must be too high, for voter turnout rarely exceeds 50 percent and is often much less. So the questionnaire method usually overestimates participation. Overcounting of voters calls into question conclusions based on the replies to these questions.[3]

As a result, the design and implementation of surveys and interviews have to be scrutinized. We begin with a thorough discussion of the problem of obtaining accurate information about attitudes and beliefs by asking people questions rather than by directly observing their behavior. This background puts us in a position to examine survey and interview methods carefully and thoroughly.

Fundamentals: Ensuring Validity and Reliability

Since survey and interview methods produce only indirect measures of attitudes and behavior, measurement problems, as discussed in chapter 5, come to the fore. In particular, what is recorded on a piece of paper or an audiotape is usually not an exact, error-free measure of an object. This is particularly true when the objects are attitudes, beliefs, or self-described behavior. An observed "score" (for example, a response to a question) is composed of a true or real (but unobserved) measure plus various types of error. The errors may be random or systematic. Random errors arise by chance or happenstance and (it is hoped) cancel one another out. A systematic error, by contrast, results when a measuring device consistently over- or underestimates a true value, as when a scale always reads two pounds less than a person's real weight. The goal of any research design, of course, is to minimize these errors. Stated differently, our investigative procedures have to ensure validity and reliability. A valid measure produces an accurate or true picture of an object, whereas a reliable one gives consistent results (measurements) across time and

[3] See Robert F. Belli, Michael W. Traugott, Margaret Young, and Katherine A. McGonagle, "Reducing Vote Overreporting in Surveys: Social Desirability, Memory Failure, and Source Monitoring," *Public Opinion Quarterly* 63, no. 1 (1999): 90–108; and Janet M. Box-Steffensmeier, Gary C. Jacobson, and J. Tobin Grant, "Question Wording and the House Vote Choice: Some Experimental Evidence," *Public Opinion Quarterly* 64, no. 1 (2000): 257–70.

users. In the case of survey research, attaining both goals can be a daunting but surmountable problem. This is an important point for making sense of claims based on polls.

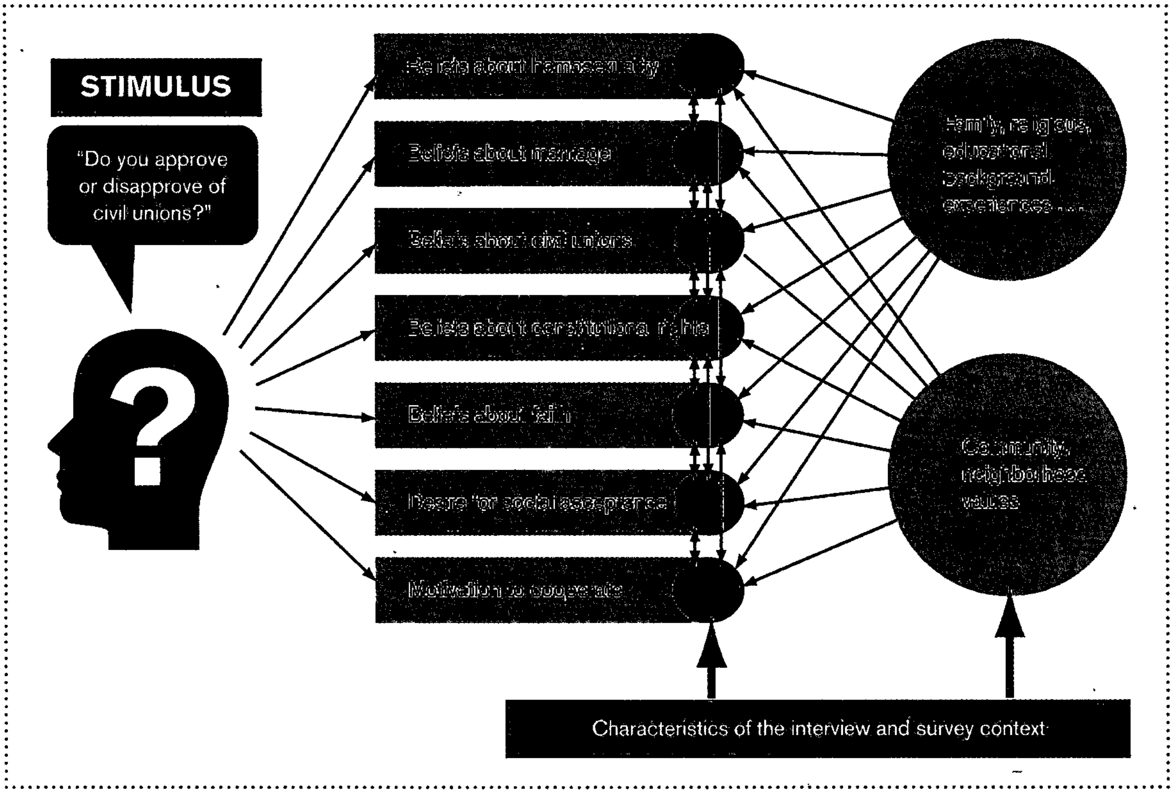
Let's think about a naive view of survey research. A pollster asks a man if he supports or opposes civil unions, a legal status that gives gay and lesbian couples rights equal to or similar to those enjoyed by traditionally married people. When asking such a question, most people expect that most of the time the responses will be a precise representation of what the respondent intends or thinks and that all parties understand the question. In the case of a supposedly objective scientific method such as survey research, the replies appear to be straightforward statements that accurately express a person's real feelings. It is usually assumed, in other words, that the response is unproblematic; that *approve* means "approve" and there is nothing more to the story than that.

To see what can go wrong, consider figure 10-1. It shows that a fully formed attitude does not simply sit in someone's mind isolated from all other mental states. Instead, a verbalized or a written opinion that is stimulated by a question is a distillation of a number of beliefs, hopes, desires, and motivations. Furthermore, one person who hears the phrase *civil union* may not be familiar with the term, whereas another, perhaps more informed about current events, may realize the question deals with marriage for homosexual couples. They may express opposing views, but the opinions can be based on unrelated beliefs about the subject. Or, if both people reply with, say, "approve," they may do so for different reasons. Even though two people may give the same response, the answers may not be comparable when this opinion is associated with other attitudes or behaviors. And consider that even if two people share a common understanding of the term, they may differ greatly in other respects: the intensity of their feelings about the matter, their willingness to cooperate with the research, or their desire to "please" the interviewer by giving a socially acceptable response. We must also factor in the interviewer's characteristics (for example, demeanor, race, gender) and the context of the research (for example, nature of the sponsoring organization, time constraints on participants). When an interviewer influences a respondent's answers through characteristics or other means it is called **interviewer bias**. The net result is that an interview, even one carried out over the phone or using a mail survey, involves a complex set of potential interactions that can confuse the interpretation of responses.

To deal with problems of this nature, even in the simplest situations, we need to ensure that several assumptions have been met, including the following (for simplicity, let *R* stand for the respondent and *I* for the interviewer):

- The requested information must be available to *R* (that is, not forgotten or misunderstood).
- *R* must know what is to *I* a relevant and appropriate response.

FIGURE 10-1 **Stimulus Activates Many Beliefs, Desires, Motives . . .**


STIMULUS
"Do you approve or disapprove of civil unions?"
?
Beliefs about homosexuality
Beliefs about marriage
Beliefs about civil unions
Beliefs about constitutional rights
Beliefs about faith
Desire for social acceptance
Motivation to cooperate
Family, religious, educational background, experiences . . .
Community, neighborhood values
Characteristics of the interview and survey context

- *R* must be motivated to provide *I* with the information.
- *R* must know how to provide the information.
- *I* must accurately record *R*'s responses.
- The responses must reflect *R*'s meanings and intentions, not *I*'s.
- Other users of the data must understand the questions and answers the same way *R* and *I* do.

The point is that if questionnaires and interviews are to produce any useful information, they must take into account the mental context of the respondent and the interview situation. For that reason, we spend the greater portion of this chapter discussing crucial topics such as question wording, questionnaire layout, administrative protocols, efforts to balance demands for completeness versus costs, ways of motivating cooperation, and interviewer characteristics and deportment. These factors contribute to the validity and reliability of the measurements or observations obtained through questionnaires and interviews.

Survey Research

The term *survey research* has two broad meanings. In the context of research design, it indicates an alternative data collection method to experiments, simulations, and formal modeling. Instead of manipulating an independent variable, for instance, to see the effect on behavior, the survey design asks people if they have been exposed to some factor and, if so, by how much, and then it relates the responses to other questions about their behavior. In this chapter, we use the term a bit more specifically to mean research based on direct or indirect interview methods. Simply stated, a group of individuals respond to or fill out more or less standardized questionnaires. (The questionnaires may take different forms to investigate different hypotheses, but they do not involve freewheeling or spontaneous conversations.) Administering questionnaires is one of the most familiar political science research methods.

As the use of surveys has grown, so too has the amount of research on the method itself. This research tries to improve the validity and reliability of the method while keeping costs manageable. We now know more about many aspects of survey research than was known when it was first used, much to the benefit of researchers and consumers of survey research.

We begin with a review of the types of surveys and some of their important characteristics. Then we take up response quality and question wording, the heart and soul of this type of research.

Types of Surveys

A survey solicits information by posing a standard set of questions and stimuli to a sample of individuals drawn from an appropriate target population (see chapter 7). The forms of the instrument and means of administration vary widely depending on a host of factors ranging from cost to comprehensiveness. Table 10-1 lists the main types of surveys along with a few of their properties or characteristics. As you can see from the table, surveys range from personal or face-to-face interviews to contacting subjects by mail or telephone to more or less hit-or-miss methods such as posting questions on a Web site or leaving them in a public area.

Perhaps the most familiar surveys are those conducted personally or face to face. The interviewer typically follows a structured questionnaire that is administered to all members of the sample, although sometimes different forms are used with slightly different question wording and order. Not only are the same questions asked of everyone, but the manner in which they are posed is also standardized to the maximum extent possible. The results are then coded or transcribed for further analysis. Moreover, for a variety of reasons, the principal investigator does not

American Association for Public Opinion Research's Transparency Initiative

Scholars and practitioners of survey research face many choices in how they conduct survey research and in how they report methods and findings. As discussed in chapter 2, the scientific method is premised on making methods and findings public. This allows other scientists and readers to evaluate the quality of the methodological approach and consider findings in the context of the methods. When it comes to survey research, consumers of polls are advantaged when they have access to important information about a poll like the sampling frame, the sample design, and the sample size.

In order to improve the information made available by survey researchers the American Association for Public Opinion Research (AAPOR) has in recent years launched a transparency initiative focused on standardizing the information researchers release to the public. According to AAPOR,

> Good professional practice imposes the obligation upon all survey and public opinion researchers to disclose certain essential information about how the research was conducted. When conducting publicly released research studies, full and complete disclosure to the public is best made at the time results are released, although some information may not be immediately available. When undertaking work for a private client, the same essential information should be made available to the client when the client is provided with the results.

AAPOR has developed a set of standards that organizations that conduct surveys for public consumption must follow if they voluntarily join the transparency initiative. These standards are designed to increase transparency of methods and allow readers to better understand how to interpret poll results. The standards appear below.

A. We shall include the following items in any report of research results or make them available immediately upon release of that report.

1. Who sponsored the research study, who conducted it, and who funded it, including, to the extent known, all original funding sources.
2. The exact wording and presentation of questions and responses whose results are reported.
3. A definition of the population under study, its geographic location, and a description of the sampling frame used to identify this population. If the sampling frame was provided by a third party, the supplier shall be named. If no frame or list was utilized, this shall be indicated.
4. A description of the sample design, giving a clear indication of the method by which the respondents were selected (or self-selected) and recruited, along with any quotas or additional sample selection criteria

applied within the survey instrument or post-fielding. The description of the sampling frame and sample design should include sufficient detail to determine whether the respondents were selected using probability or non-probability methods.

5. Sample sizes and a discussion of the precision of the findings, including estimates of sampling error for probability samples and a description of the variables used in any weighting or estimating procedures. The discussion of the precision of the findings should state whether or not the reported margins of sampling error or statistical analyses have been adjusted for the design effect due to clustering and weighting, if any.
6. Which results are based on parts of the sample, rather than on the total sample, and the size of such parts.
7. Method and dates of data collection.

B. We shall make the following items available within 30 days of any request for such materials.

1. Preceding interviewer or respondent instructions and any preceding questions or instructions that might reasonably be expected to influence responses to the reported results.
2. Any relevant stimuli, such as visual or sensory exhibits or show cards.
3. A description of the sampling frame's coverage of the target population.
4. The methods used to recruit the panel, if the sample was drawn from a pre-recruited panel or pool of respondents.
5. Details about the sample design, including eligibility for participation, screening procedures, the nature of any oversamples, and compensation/incentives offered (if any).
6. Summaries of the disposition of study-specific sample records so that response rates for probability samples and participation rates for non-probability samples can be computed.
7. Sources of weighting parameters and method by which weights are applied.
8. Procedures undertaken to verify data. Where applicable, methods of interviewer training, supervision, and monitoring shall also be disclosed.

C. If response rates are reported, response rates should be computed according to AAPOR Standard Definitions.

D. If the results reported are based on multiple samples or multiple modes, the preceding items shall be disclosed for each.

E. If any of our work becomes the subject of a formal investigation of an alleged violation of this *Code*, undertaken with the approval of the AAPOR Executive Council, we shall provide additional information on the research study in such detail that a fellow researcher would be able to conduct a professional evaluation of the study.

Source: American Association for Public Opinion Research, AAPOR Code of Ethics, revised May 2010, section III Standards for Disclosure. Available at www.aapor.org/AAPORKentico/Standards-Ethics/AAPOR-Code-of-Ethics.aspx

TABLE 10-1 Types and Characteristics of Surveys

Type of Survey	Overall Cost[a]	Potential Completion Rate[b]	Characteristics Sample-Population Congruence	Questionnaire Length[c]	Data-Processing Costs
Personal/face to face	High	High to medium	Potentially high	Long-medium	High
Telephone	Medium	Medium	Medium	Medium-short	High to low
Mail	Low	Low	Medium	Medium-short	Medium
E-mail	Low	Depends but low	Low	Medium-short	High to low
Internet	Low	Depends but low	Low	Medium-short	High to low
Group administration	Very low	High once group is convened	Depends on group selection process	Variable	High to low
Drop-off/pick-up	Very low	Low	Low	Short	Low

[a] Costs of design, administration, and processing.

[b] Assumes a general target population (see text): high = greater than 75 percent; medium = 30 to 75 percent; low = less than 30 percent.

[c] *Length* can refer to the number of questions or the time to complete (see text).

usually conduct the interviews but uses paid or volunteer assistants. Hence, this kind of research can be quite expensive.[4]

Academic and commercial polls are increasingly being conducted in whole or part by mail, phone, or the Internet. A mail survey, which may be preceded by an introductory letter, has to be self-contained with clear questions and instructions for completing and returning it. Motivating participants, anticipating misunderstandings, and obtaining unambiguous results demand a lot of careful planning and pretesting. They also require a list of addresses drawn from the population of interest. Although somewhat less expensive, phone interviews raise a number of tricky problems of their own (discussed below). Nevertheless, the basic idea is the same: pose a series of questions or stimuli and record the responses.

Internet surveys are an increasingly important type of survey. Web sites like YouGov have capitalized on the growing market for Internet surveys and have carved out an important presence in polling the American public.[5] Internet surveys have become

[4] Indeed, a large (1,000 or more respondents) national survey using probability sampling of the sort explained in chapter 7 might cost more than $100,000.

[5] YouGov, https://today.yougov.com/#/

increasingly popular in part because of some important advantages over other types of surveys including lower costs, the ability to question respondents about a wide variety of multimedia materials like pictures or video segments and allow respondents to answer questions when it is most convenient and at their own pace. But there are some significant issues associated with Internet surveys that make them both interesting examples and vexing to survey researchers.

The biggest obstacle associated with Internet surveys is twofold. First, while Internet access has grown considerably over the last two decades, from an estimated 14 percent of American adults in 1995 to 87 percent in 2014, the Internet is still not as widely adopted as the telephone.[6] That is a critical problem when trying to create a representative sample because Internet use is correlated with important demographic indicators like age (97 percent of those aged seventeen to twenty-nine use the Internet, but only 57 percent of those aged sixty-five and older use the Internet), income (99 percent of those with family incomes over $75,000 use the Internet, but only 77 percent of those with family incomes under $30,000 use the Internet), and education (97 percent of college graduates use the Internet, while only 76 percent of those who did not attend college do so).[7] Simply put, not everyone uses the Internet, and if it were possible to randomly sample all Internet users, that sample would not be representative of the US population. Second, unlike telephone or mail surveys, where it is possible to assemble nearly complete lists of phone numbers or mailing addresses for identifying a population from which to sample, there is no comparable way to identify all Internet users. Internet surveys are therefore limited to those that use select lists of e-mail addresses or social media accounts, or those that collect responses from respondents who happen to see the survey on a Web site.

For example, consider *Time* magazine's Internet survey to help determine its Person of the Century in 1999. *Time* collected survey responses from readers on its Web site to help make the decision. At one point during the survey, *Time*'s survey results listed Mustafa Kemal Ataturk as the top vote recipient. Ataturk, who is credited with founding the Republic of Turkey in 1923, received overwhelmingly popular support after Turkish newspapers encouraged readers to vote for Ataturk to be *Time*'s Person of the Century. In the end, *Time* chose Albert Einstein instead.[8] The *Time* example highlights the risks researchers take when relying on Internet surveys that collect responses from users without any sampling or selection process.

To overcome this obstacle, many survey researchers who want to use an Internet survey because of its many advantages choose to identify a random sample by first

[6] Pew Research Center for the People & the Press. Accessed February 10, 2015. Available at http://www.pewinternet.org/data-trend/internet-use/internet-use-over-time/

[7] Ibid. Available at http://www.pewinternet.org/data-trend/internet-use/latest-stats/

[8] American Association of Public Opinion Research. Accessed February 10, 2015. Available at http://www.aapor.org/AAPORKentico/Education-Resources/For-Researchers/Poll-Survey-FAQ/Bad-Samples.aspx

contacting respondents via phone or mail. Respondents are asked to indicate if they have Internet access, and a sample is then identified from these respondents. Many Internet surveys could therefore be appropriately categorized as hybrid surveys because they make use of different survey methods to identify respondents and collect data. Other researchers rely on panels of respondents who answer multiple surveys on the Internet over time. YouGov relies on panels of respondents who participate in multiple surveys enticed by participation points that can be exchanged for prizes.[9]

Finally, it is possible, even necessary sometimes, to prepare a survey that is administered to a group (for example, a political science class or visitors to a senior center) or made available at a public location (a library, museum, or dormitory lounge). The finished forms are then collected or returned to the same or another convenient spot. The results are generally suspect in some people's minds and probably not publishable because they may not be representative, but the method offers considerable savings in effort and cost. For this reason they are commonly used at schools and colleges.

As might be expected, each of these types has advantages and disadvantages. The entries in table 10-1 are merely suggestive and comparative, and have to be interpreted flexibly. A phone survey, for example, generally has to be shorter (in time necessary to complete it) than a personal interview because respondents to the former may be reluctant to tie up a phone line or may be distracted by those around them. Given an interesting topic, plenty of forewarning, and trained interviewers, however, it is possible to hold people's attention for longer periods.

Characteristics of Surveys

COST. Any type of survey research takes time and incurs at least some expenses for materials. Among the factors determining survey costs are the amount of professional time required for questionnaire design, the length of the questionnaire, the geographical dispersion of the sample, callback procedures, respondent selection rules, and availability of trained staff.[10] Personal interviews are the most expensive to conduct because interviewers must, after being trained, locate and contact subjects, a frequently time-consuming process. For example, some well-established surveys ask interviewers to visit a household and, if the designated subject is not available, make one or more callbacks. National in-person surveys also incur greater administrative costs. Regional supervisory personnel must be hired and survey instruments sent back and forth between the researcher and the interviewers. Mail surveys are less expensive but require postage and materials. Electronic surveys (for example,

[9] YouGov. Accessed February 10, 2015. Available at https://today.yougov.com/account/login/?next=%2Fopi%2Fmyfeed

[10] Floyd J. Fowler, *Survey Research Methods*, rev. ed. (Newbury Park, Calif.: Sage, 1988), 68.

e-mail or Internet) do not necessitate interviewer time but must still be set up by individuals with technical skills. Although mail surveys are thought to be less expensive than telephone polls, Fowler argued that the cost of properly executed mail surveys is likely to be similar to that of phone surveys.[11] Thus, when deciding among personal interviews, telephone interviews, and mail surveys, researchers must consider cost and administrative issues.

Compared with personal interviewing, telephone surveys have several administrative advantages.[12] Despite the cost of long-distance calls, centralization of survey administration means that training of telephone interviewers may be easier, and flexible working hours are often attractive to the employees. But the real advantages to telephone surveys begin after interviewing starts. It is possible to exercise greater supervision over interviewers and give them prompt feedback. Also, callers can easily inform researchers of any problems they encounter with the survey. Coders can begin coding data immediately. If they discover any errors, they can inform interviewers before a large problem emerges. With proper facilities, interviewers may be able to code respondents' answers directly on computer terminals. In some cases, the whole interview schedule may be computerized, with questions and responses displayed on a screen in front of the interviewer. These are known as computer-assisted telephone interviews (CATIs). Computer and telephone technologies give telephone surveys a significant time advantage over personal interviews and mail surveys. Telephone interviews may be completed and data analyzed almost immediately.[13] On the downside, people tend to be home mostly in the evenings and weekends, but calls made during these hours often meet resistance. This problem used to be aggravated by an explosion in the use of telemarketing; now, however, with the advent of "do-not-call" lists, the situation may not be as dire as it once was.

Almost needless to say, group surveys (those that are distributed to members of groups who might be expected to fill them out at a group meeting or online) and drop-off surveys (questionnaires that are left in public places like libraries, malls, or offices with collection boxes on-site) are least expensive. After all, they require minimal administrative and personnel costs to gather the data. On the other hand, the questionnaires still have to be carefully constructed and tested. This is particularly true if the investigator is not in a position to provide guidance or answer questions during survey administration.

Internet surveys still face the same costs associated with developing survey instruments. Researchers can realize tremendous cost savings through the use of Internet surveys. Internet surveys reduce postal, telephone, and labor costs because

[11] Ibid.

[12] Robert M. Groves and Robert L. Kahn, *Surveys by Telephone: A National Comparison with Personal Interviews* (New York: Academic Press, 1979); and James H. Frey, *Survey Research by Telephone* (Beverly Hills, Calif.: Sage, 1983).

[13] Frey, *Survey Research by Telephone*, 24–25.

respondents can answer surveys at any time of day, from any part of the world, by replying to an e-mail or answering questions online. Not only does the increase in broadband and Internet-enabled smartphones and tablets make the use of video conferencing services like Skype feasible survey research tools, it can also dramatically reduce the costs of in-person interviews because researchers can communicate with respondents around the globe, face to face, without incurring expensive travel costs.

COMPLETION RATES. One of the maddening characteristics of many commercial polls is that they often do not indicate how many people refused to take part in the survey. As a typical example, a CBS poll contained the following information: "This poll was conducted by telephone on August 2–3, 2011 among 960 adults nationwide."[14] This number, however, most likely refers to the number of complete or nearly complete questionnaires and not to the refusals to participate in the survey in the first place. This information can be important to have.

A completion rate or **response rate** refers to the proportion of persons initially contacted who actually participate. In a mail survey, for instance, the denominator is the total number of questionnaires sent out, not the number returned. Three distinguished researchers, Robert M. Groves, Robert B. Cialdini, and Mick P. Couper, succinctly summarized the significance of this quantity for the social sciences:

> Among the alternative means of gathering information about society, survey research methods offer unique inferential power to known populations. . . . This power, however, is the cumulative result of many individual decisions of sample persons to participate in the survey. When full participation fails to obtain, the inferential value of the method is threatened.[15]

We need to explore this point in slightly greater detail. If the response rate is low, either because individuals cannot be reached or because they refuse to participate, the researchers' ability to make statistical inferences for the population being studied may be limited. Also, those who do participate may differ systematically from those who do not, creating other biases. Increasing the size of the survey sample to compensate for low response rates may only increase costs without alleviating the problem.

Most of what we know about response rates comes from studies of personal interview, mail, and telephone surveys. It is difficult, perhaps impossible in some cases,

[14] "Poll: Disapproval of Congress Hits All-Time High," *CBS News Political Hotsheet* (blog), August 4, 2011. Available at http://www.cbsnews.com/8301-503544_162-20088388-503544.html

[15] Robert M. Groves, Robert B. Cialdini, and Mick P. Couper, "Understanding the Decision to Participate," *Public Opinion Quarterly* 56, no. 4 (1992): 474.

to measure participation levels in electronic and drop-off studies. At one time, response rates were clearly superior for personal interview surveys of the general population than for other types of surveys. Response rates of 80 to 85 percent were often required for federally funded surveys.[16] Higher response rates were not uncommon. By the 1970s, however, response rates for personal interview surveys declined. In 1979 it was reported that in "the central cities of large metropolitan areas the final proportion of respondents that are located *and* consent to an interview is declining to a rate sometimes close to 50 percent."[17]

In general, the decrease in response rates for personal interview surveys has been attributed to both an increased difficulty in contacting respondents and an increased reluctance among the population to participate in surveys. There are more households now in which all adults work outside the home, which makes it difficult for interviewers to get responses. Moreover, pollsters continually worry about public resistance to their craft.[18]

In large cities, nonresponse can be attributed to several additional factors: respondents are less likely to be at home, are more likely not to have a full command of English, or both; interviewers are less likely to enter certain neighborhoods after dark; and security arrangements in multiple-unit apartment buildings make it difficult for interviewers to reach potential respondents. Moreover, many individuals such as undocumented immigrants or people receiving welfare benefits are often skittish about talking to "official-looking" strangers. Because of poor working conditions, it is hard to find skilled and experienced interviewers to work in large cities. In smaller cities and towns as well, people have shown an increased tendency to refuse to participate in surveys.[19]

Higher refusal rates may be due to greater distrust of strangers and fear of crime as well as to the increased number of polls. For example, in one study of respondents' attitudes toward surveys, about one-third did not believe that survey participation benefited the respondent or influenced government.[20] An equal number thought that too many surveys were conducted and that too many personal questions were asked. Some survey researchers feared that the National Privacy Act, which requires researchers to inform respondents that their participation is voluntary, would lead

[16] Earl R. Babbie, *Survey Research Methods* (Belmont, Calif.: Wadsworth, 1973), 171.

[17] Groves and Kahn, *Surveys by Telephone*, 3.

[18] See, for example, Burns W. Roper, "Evaluating Polls with Poll Data," *Public Opinion Quarterly* 50, no. 1 (1986): 10–16.

[19] Charlotte G. Steeh, "Trends in Nonresponse Rates, 1952–1979," *Public Opinion Quarterly* 45, no. 1 (1981): 40–57.

[20] Laure M. Sharp and Joanne Frankel, "Respondent Burden: A Test of Some Common Assumptions," *Public Opinion Quarterly* 47, no. 1 (1983): 36–53. Note that another survey found that people had generally favorable beliefs about polls (Roper, "Evaluating Polls with Poll Data"), but even the author of this study worried that the public might grow weary and distrustful of polling.

to more refusals. However, one study found that privacy concerns and past survey experience were more frequent reasons for refusal than was being informed of the voluntary nature of participation.[21]

Some of these findings about why people do not participate in personal interview surveys raise the possibility that survey research of all types may become increasingly difficult to conduct. The increased nonresponse has reduced the advantage of the personal interview over mail and telephone surveys. In fact, Dillman, using his "total design method" for mail and telephone surveys, has achieved response rates rivaling those for personal interviews.[22] He concluded that the chance someone will agree to be surveyed is best for the personal interview but that telephone interviews are now a close second, followed by mail surveys. Other research comparing response rates of telephone and personal interview surveys has also found little difference.[23]

Two norms of telephone usage have contributed to success in contacting respondents by phone and completing telephone interviews.[24] First, most people feel compelled to answer the phone if they are home when it rings. A telephone call represents the potential for a positive social exchange. With the increase in telephone solicitation and surveys, this norm may be revised, however. Caller ID and answering machines can be used to screen and redirect unwanted calls. Thus, telephone surveys may increasingly become prearranged and conducted after contact has been established by some other method.

A second norm of telephone usage is that the initiator should terminate the call. This norm gives the interviewer the opportunity to introduce himself or herself. And in a telephone interview the introductory statement is crucial (see the following discussion on motivation). Because the respondent lacks any visual cues about the caller, the initial response is one of uncertainty and distrust. Unless the caller can quickly alleviate the respondent's discomfort, the respondent may refuse

[21] Theresa J. DeMaio, "Refusals: Who, Where, and Why," *Public Opinion Quarterly* 44, no. 2 (1980): 223–33.

[22] Dillman, *Mail and Telephone Surveys.*

[23] See Theresa F. Rogers, "Interviews by Telephone and in Person: Quality of Responses and Field Performance," *Public Opinion Quarterly* 40, no. 1 (1976): 51–65; and Groves and Kahn, *Surveys by Telephone.* Response rates are affected by different methods of calculating rates for the three types of surveys. For example, nonreachable and ineligible persons may be dropped from the total survey population for telephone and personal interviews before response rates are calculated. Response rates to mail surveys are depressed because all nonresponses are assumed to be refusals, not ineligibles or nonreachables. Telephone response rates may be depressed if nonworking but ringing numbers are treated as nonreachable but eligible respondents. Telephone companies vary in their willingness to identify working numbers. If noneligibility is likely to be a problem in a mail survey, ineligibles should be asked to return the questionnaire anyway, so that they can be identified and distinguished from refusals.

[24] Frey, *Survey Research by Telephone,* 15–16.

to finish the interview. For this reason, telephone interviews are more likely to be terminated before completion than are personal interviews. It is harder to ask an interviewer to leave than it is simply to hang up the phone.

Because of the importance attached to high response rates, much research on how to achieve them has been conducted. For example, an introductory letter sent prior to a telephone interview has been found to reduce refusal rates.[25] In fact, such letters may result in response rates that do not differ significantly from those for personal surveys.[26] Researchers have also investigated the best times to find people at home. One study found that for telephone interviews, evening hours are best (6:00 to 6:59, especially), with little variation by day (weekends excluded).[27] Another study concluded that the best times for finding someone at home were late afternoon and early evening during weekdays, although Saturday until four in the afternoon was the best time overall.[28]

Because mail surveys usually have the poorest response rates, many researchers have investigated ways to increase responses to them.[29] Incentives (money, pens, and other token gifts) have been found to be effective, and prepaid incentives are better than promised incentives. Follow-up, prior contact, type of postage, sponsorship, and title of the person who signs the accompanying letter are also important factors in improving response rates. Telephone calls made prior to mailing a survey may increase response rates by alerting respondents to the survey's arrival. Telephone calls also are a quick method of reminding respondents to complete

[25] Don A. Dillman, Jean Gorton Gallegos, and James H. Frey, "Reducing Refusal Rates for Telephone Interviews," *Public Opinion Quarterly* 40, no. 1 (1976): 66–78.

[26] Fowler, *Survey Research Methods*, 67.

[27] Gideon Vigderhous, "Scheduling Telephone Interviews: A Study of Seasonal Patterns," *Public Opinion Quarterly* 45, no. 2 (1981): 250–59.

[28] Michael F. Weeks, Bruce L. Jones, R. E. Folsom, and Charles H. Benrud, "Optimal Times to Contact Sample Households," *Public Opinion Quarterly* 44, no. 1 (1980): 101–14.

[29] See J. Scott Armstrong, "Monetary Incentive in Mail Surveys," *Public Opinion Quarterly* 39 (1975): 111–16; Arnold S. Linsky, "Stimulating Responses to Mailed Questionnaires: A Review," *Public Opinion Quarterly* 39, no. 1 (1975): 82–101; James R. Chromy and Daniel G. Horvitz, "The Use of Monetary Incentives in National Assessment Household Surveys," *Journal of the American Statistical Association* 73 (1978): 473–78; Thomas A. Heberlein and Robert Baumgartner, "Factors Affecting Response Rates to Mailed Questionnaires: A Quantitative Analysis of the Published Literature," *American Sociological Review* 43, no. 4 (1978): 447–62; R. Kenneth Godwin, "The Consequences of Large Monetary Incentives in Mail Surveys of Elites," *Public Opinion Quarterly* 43, no. 3 (1979): 378–87; Kent L. Tedin and C. Richard Hofstetfer, "The Effect of Cost and Importance Factors on the Return Rate for Single and Multiple Mailings," *Public Opinion Quarterly* 46, no. 1 (1982): 122–28; Anton J. Nederhof, "The Effects of Material Incentives in Mail Surveys: Two Studies," *Public Opinion Quarterly* 47, no. 1 (1983): 103–11; Charles D. Schewe and Norman G. Cournoyer, "Prepaid vs. Promised Monetary Incentives to Questionnaire Response: Further Evidence," *Public Opinion Quarterly* 40, no. 1 (1976): 105–07; James R. Henley Jr., "Response Rate to Mail Questionnaire with a Return Deadline," *Public Opinion Quarterly* 40 (1976): 374–75; Thomas A. Heberlein and Robert Baumgartner, "Is a Questionnaire Necessary in a Second Mailing?" *Public Opinion Quarterly* 45, no. 1 (1981): 102–08; and Wesley H. Jones, "Generalizing Mail Survey Inducement Methods: Population Interactions with Anonymity and Sponsorship," *Public Opinion Quarterly* 43, no. 1 (1979): 102–11.

and return questionnaires. Good follow-up procedures allow a researcher to distinguish between respondents who have replied and those who have not without violating the anonymity of respondents' answers. Generally, mail surveys work best when the population is highly literate and interested in the research topic under investigation.[30]

Internet surveys represent a double-edged sword with respect to response and completion rates. On the one hand, Internet surveys can be answered at the respondent's convenience, which improves completion rates compared to other survey types. Also, a respondent is more likely to simply press "submit" at the end of a survey than to seal an envelope and return a mail survey, which improves response rates over a mail survey. On the other hand, requests to participate in Internet surveys can also be more easily ignored than a researcher physically knocking on the door or calling on the phone. Who among us has not simply deleted an e-mail request from an unknown sender without a second thought?

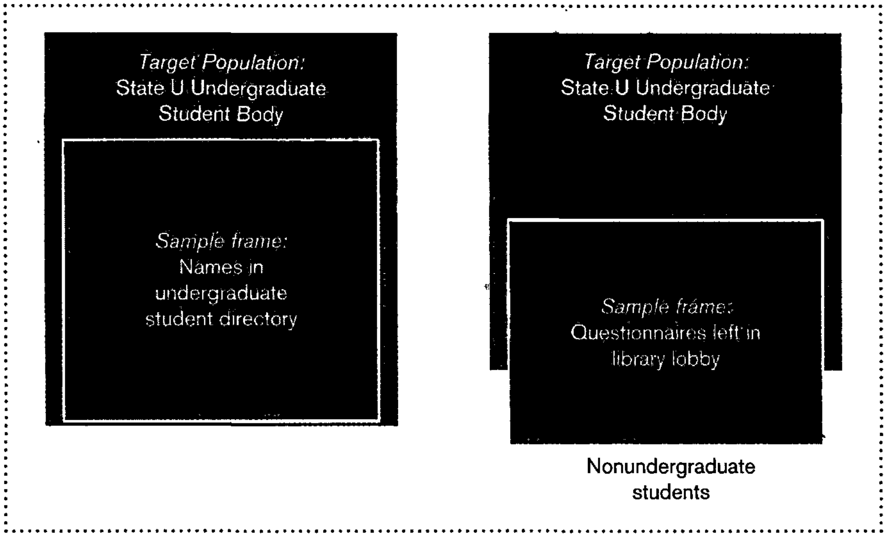
In sum, response rates are an important consideration in survey research. When evaluating research findings based on survey research, you should check the response rate and what measures, if any, were taken to increase it. Should you ever conduct a survey of your own, a wealth of information is available to help you to achieve adequate response rates.

SAMPLE-POPULATION CONGRUENCE. **Sample-population congruence**, which refers to how well the sample subjects represent the population, is always a major concern. Here we are speaking of how well the individuals in a sample represent the population from which they are presumably drawn. Bias can enter either through the initial selection of respondents or through incomplete responses of those who agree to take part in the study. In either case a mismatch exists between the sample and the population of interest. These problems arise to varying degrees in every type of survey.

Some of the cheapest and easiest surveys, such as drop-off or group questionnaires, encounter difficulties in matching sampling frames with the target population, as figure 10-2 suggests. Suppose, for example, you wanted to survey undergraduates at your college about abortion. One option would be to draw a sample of names and addresses from the student directory. Assuming all currently enrolled students are correctly listed there, the sampling frame (the directory) should closely match the target population, the undergraduate student body. A random sample drawn from the list would presumably be representative. If instead you left a pile of questionnaires in the library, you would have much less

[30] Fowler, *Survey Research Methods*, 63.

FIGURE 10-2 **Matching Sampling Frames to Target Population**



control over who responds. Now your "sample" might include graduate students, staff, and outside visitors. It would then be difficult to draw inferences about the student body. One solution would be to add a "filter" question (e.g., "Are you a freshman, sophomore, . . . ?") to sort out nonstudents, but the potential for problems can easily be imagined.[31]

Recall from chapter 7 that when all members of a population are listed, theoretically there is an equal opportunity for all members to be included in the sample. Only rarely, however, are all members included. Personal interviews based on cluster samples in which all the households of the last sampled cluster or area are enumerated and then selected at random, which gives each household an equal chance of being selected, are likely to be more representative than are mail or telephone surveys based on published lists, which are often incomplete.

Telephone surveys attempt to improve the representativeness of samples with a procedure called random digit dialing (the use of randomly generated telephone numbers instead of telephone directories; see chapter 7) and by correcting for households with more than one number. Thus, people who have unlisted numbers or new numbers may be included in the sample. Otherwise, a telephone survey may be biased by the exclusion of these households. Estimates of the number of households in the United States that do not have phones vary from 2 to 10 percent,

[31] And, of course, you still would not have a probability sample. See chapter 7.

whereas only about 5 percent of dwelling units are missed with personal interview sampling procedures.[32]

A relatively recent technical innovation has put a wrinkle into telephone interviewing: the rise of cell-phone-only users. During the competitive 2004 presidential election, polling companies and media outlets worried that unless this growing group could be included in their sampling frames, predictions about the outcome could be in error. That is, if interest lies in surveying the general public, and a small but noticeable fraction of people are excluded from the sample frame, bias might creep into the study. Apparently, there was some basis for concern. According to a Pew Research Center report, in 2006 about 7 to 9 percent of Americans relied on cell phones instead of landlines or combinations of cell and landline phones. More ominously, this subpopulation differs from the general public in several significant ways. Cell-phone-only individuals tend to be younger, earn less, and hold somewhat more "liberal" views on social issues (for example, abortion, gay marriage) compared with the populace as a whole.[33] The study also pointed out that cell-phone-only users are easier to contact than landline users, but doing so is more difficult and expensive, and that response rates are lower and refusal rates are higher among cell-phone-only users. Nevertheless, by making statistical adjustments, the Pew Research Center found that substantive conclusions about attitudes on political issues were not greatly distorted.

During the 2012 presidential election, however, multiple polling organizations' efforts at estimating the support for Democrat Barack Obama and Republican Mitt Romney were stymied by cell-phone-only voters. By the 2012 election, one-third of Americans were cell-phone-only. Some organizations that included cell phones in telephone surveys, like IBD/TIPP, generated poll results in the days leading up to Election Day that closely mirrored the eventual national presidential vote. Other organizations that included cell phones in telephone polling, like Gallup, were off the mark, predicting a Romney victory.[34] Polling organizations continue to struggle with how to appropriately capture cell-phone-only voters.

Internet surveys, as discussed earlier in the chapter, are even more prone to a fault in sample-population congruence. While phone books offer incomplete lists of phone numbers and addresses, the lists can be used as a starting point for identifying a complete sampling frame. Identifying a population through the Internet is a

[32] Groves and Kahn, *Surveys by Telephone*, 214; and Frey, *Survey Research by Telephone*, 22.

[33] Pew Research Center for People & the Press, "The Cell Phone Challenge to Survey Research," May 15, 2006. http://people-press.org/2006/05/15/the-cell-phone-challenge-to-survey-research/

[34] Nate Silver, "Which Polls Fared Best (and Worst) in the 2012 Presidential Race," *New York Times*, November 10, 2012. Available at http://fivethirtyeight.blogs.nytimes.com/2012/11/10/which-polls-fared-best-and-worst-in-the-2012-presidential-race/?_r=0

fool's errand in most cases, as directories of addresses simply do not exist for most populations of interest.

As mentioned earlier, one of the major reasons for worry about sample-population congruence is the possibility that those who are not included will differ from those who are.[35] Some evidence of this likelihood appears in the literature. African Americans, for example, have been found to be more likely to refuse telephone interviews.[36] Refusals also are more common among older, middle-class persons; urban residents; and westerners.[37]

The amount of bias introduced by nonresponses due to refusal or unavailability varies, depending on the purpose of the study and the explanatory factors stressed by the research. For example, if urbanization was a key explanatory variable and refusals were concentrated in urban areas, the study could misrepresent respondents from urban areas because the urban respondents who agreed to participate could differ systematically from those who refused. The personal interview provides the best direct opportunity to judge the characteristics of those who refuse and to estimate whether their refusals will bias the analysis.[38]

The bottom line is that you should always ascertain how well sample proportions match the population of interest on key demographic variables (e.g., age, gender, ethnicity). Nearly every survey analyst takes this first step, even if the results are not always reported.

QUESTIONNAIRE LENGTH. Subject fatigue is always a problem in survey research. Thus, if a survey poses an inordinate number of questions or takes up too much of the respondents' time, the respondents may lose interest or start answering without much thought or care. Or they may get distracted, impatient, or even hostile. And keeping people interested in the research is exactly what is needed.

[35] For research estimating the amount of bias introduced by nonresponse due to unavailability or refusal, see F. L. Filion, "Estimating Bias Due to Nonresponse in Mail Surveys," *Public Opinion Quarterly* 39, no. 4 (1975–96): 482–92; Michael J. O'Neil, "Estimating the Nonresponse Bias Due to Refusals in Telephone Surveys," *Public Opinion Quarterly* 43, no. 2 (1979): 218–32; and Arthur L. Stinchcombe, Calvin Jones, and Paul Sheatsley, "Nonresponse Bias for Attitude Questions," *Public Opinion Quarterly* 45, no. 3 (1981): 359–75.

[36] Carol S. Aneshensel, Ralph R. Frerichs, Virginia A. Clark, and Patricia A. Yokopenic, "Measuring Depression in the Community: A Comparison of Telephone and Personal Interviews," *Public Opinion Quarterly* 46, no. 1 (1982): 110–21.

[37] DeMaio, "Refusals: Who, Where, and Why," 223–33; and Steeh, "Trends in Nonresponse Rates," 40–57.

[38] Dillman, *Mail and Telephone Surveys.*

HELPFUL HINTS

Is the Sample Good Enough?

Survey analysts commonly compare their sample results with known population quantities. They do so by asking a few questions about demographic characteristics such as gender, age, and residency (e.g., urban versus rural). If these quantities are known for the target population, the sample results can be compared with the known quantities and any discrepancies noted.

Suppose, for example, that you wanted to survey undergraduates at a college but had limited resources and had to rely on a relatively low-cost class or drop-off procedure (see text). You might ask students in several large introductory courses to participate in your study. Imagine that your questionnaire asks about gender, class (e.g., freshman, sophomore), residency (in-state or out-of-state), and living arrangements (e.g., dorm, off-campus apartment). The registrar or office of institutional research probably makes these data available online or in print. Finding them should not be a major problem. Compare the sample percentages with those for the entire school. If the sample percentages are low or high in a category, you can factor that into your analysis. This method works for any group or population for which adequate measures of common traits are available.

Check out more Helpful Hints at **edge.sagepub.com/johnson8e**

A survey needs to include enough questions to gather the data necessary for the research topic, but a good rule of thumb is to keep the survey as short as possible. Given that general advice, how many questions are too many? It is almost impossible to give a precise answer. This is why hypotheses have to be stated carefully. A fishing expedition will in all likelihood end up producing little useful information. Getting the length right is another reason why repeated pretesting is crucial. By trying out the questionnaire on the same types of subjects who will be included in the final study, it is possible to head off respondent weariness.

Unless the pool of participants is especially well motivated (see below), a couple of dozen items may be the limit. Especially when there is only a limited possibility to interact with subjects (e.g., with mail, Internet and e-mail, or drop-off surveys), the number of questions should be kept to a minimum (within the confines of the project's goals, of course). Alternatively, the questionnaire should take less than, say, forty minutes (and, for phone surveys, much less time). This seems to be about the time needed for surveys conducted by government and academic institutions. A respondent's attention may be held longer in some situations than in others; hence, some questionnaires can be longer. Personal interviews generally permit researchers to ask more questions than do phone or mail surveys, and certainly more than can be asked on dropped-off forms, especially if experienced and personable interviewers do the interviewing. But again, researchers need to experiment before going into the field.

DATA-PROCESSING ISSUES. Finally, although technology is making data collection, preparation, and analysis easier, data processing remains an important subject. After the surveys have been collected, the answers still have to be tabulated. And that requirement can be costly. Consider a written questionnaire administered to five hundred people that contains fifty agree-disagree items plus ten other questions for a total of sixty. The responses will simply be marks on pieces of paper. These data need to be coded (translated) in such a way that a computer can process them. (Usually an "agree" answer would be coded as, say, a 1 and a "disagree" as a 5.) If all of the responses can be given numeric codes, there will be 60 × 500 or 30,000 bits of data to record. If, however, any of the questions are open-ended, with respondents replying in their own words, this information has to be transcribed and coded. The task used to be done laboriously by hand. If proper forms have been used, scanners can be put to work. Otherwise, the numbers or codes still have to be entered manually. In the days of IBM cards and keypunches, these chores were the bane of the survey researcher. Since the late 1990s, software has become available for this purpose, although it can be expensive and requires training to use.[1] And, as might be expected, skeptics wonder if machines can ever really decode verbatim transcripts.

Data-processing costs are a major reason for the adoption of Internet and even telephone surveys (CATIs). In the latter case, an operator uses a monitor and software to guide the interviewer through the questionnaire and record the data. One company puts it this way:

> The most important aspect of a CATI system is that it uses computers to conduct the interviews. Because a computer controls the questionnaire, skip patterns are executed exactly as intended, responses are within range, and there are no missing data. And, because answers are entered directly into the computer, data entry is eliminated—data analysis can start immediately.[2]

[1] See, for example, Matthew B. Miles and A. Michael Huberman, *Qualitative Data Analysis: An Expanded Sourcebook,* 2nd ed. (Thousand Oaks, Calif.: Sage, 1995); and Renata Tesch, *Qualitative Research: Analysis Types and Software Tools* (Bristol, Penn.: The Falmer Press, 1990).

[2] SawTooth Technologies, "WinCati for Computer-Assisted Telephone Interviewing." Accessed March 11, 2007. Available at http://www.sawtooth.com/

The software also dials the numbers, records no-answers, and handles many other administrative details. Internet surveys represent a best-case scenario when it comes to data processing as respondents process the data themselves by typing numerical responses, selecting radio buttons or sliding sliders on scales. The answers are then recorded in a data table ready for analysis. Open-ended questions must still be coded, but there is software available to analyze digitized open-ended answers (see chapter 9).

Response Quality

As we said at the outset, it is easy to take respondents' answers at face value. But doing so can be a mistake. A mere checkmark next to "approve" may or may not indicate a person's true feelings. Also, political scientists and other specialists often forget that not everyone shares their enthusiasm for or knowledge of current events. What is exciting to one person may bore another. If you are a political science major, terms and names like *party identification, World Trade Organization, Senate filibuster, ISIL (ISIS),* and *Senator Mitch McConnell* may have a clear meaning. But the public as a whole may not be nearly as familiar with them. Or they may be aware of issues such as *global warming* and *greenhouse gases* but not comprehend *volatile organic compounds* or *CO_2 emissions.* Nor will they always understand a question the way you do. You may know that *Roe v. Wade* invalidated state laws outlawing abortion, but a person in the street may only be aware that a controversial Supreme Court decision "legalized" abortion. Asking about *Roe v. Wade* might produce too many "don't know" or "no opinion" responses. Equally important, people may be reluctant to express their opinions to strangers, even if they can do so anonymously, or they may view social research as trivial or a waste of time. Finally, everyone seems to be busy; what is in your mind a short interview may be a major interference in someone else's busy life. All of these factors may affect the quality of the data obtained through a survey or an interview.

These observations lead to some important guidelines. You can apply them in your own research and, more important, should be on the lookout for how others handle (or do not handle) them.

- Motivate respondents. Good survey researchers try hard not just to induce people to take part in their studies but also to do so as enthusiastically as possible. They want more than perfunctory responses; they hope participants will be careful and thoughtful.
- Always pretest a questionnaire with the types of respondents to be included in the study, not just your friends or colleagues. Find out ahead of time what works and what doesn't.
- Be neat, organized, professional, and courteous.

- If you are using interviewers, train them especially in the skill and art of putting subjects at ease and probing and clarifying. The more experience they have, the better. Make sure they don't betray any political, ethnic, gender, age, class, or other biases that would affect the truthfulness of responses.
- Have reasonable expectations. It is not possible to conduct "the perfect study." As desirable as a personal or mail survey may be, it may not be feasible. So think about adopting an alternative and making it as rigorous as your resources allow. Regardless of the choice, keep in mind that some types of surveys have advantages over others in regard to response quality.

These guidelines pertain to **response quality**, which refers to the extent to which responses provide accurate and complete information. It is the key to making valid inferences. Response quality depends on several factors, including the respondents' motivations, their ability to understand and follow directions, their relationship with the interviewer and sponsoring organization, and, most important, the quality of the questions being asked. Indeed, this last point is so important that we discuss it in a separate section.

ENGAGING RESPONDENTS. To engage respondents, it is important to get off on a good footing by introducing yourself, your organization, your purpose, your appreciation of their time and trouble, your nonpartisanship, your awareness of the importance of anonymity, and your willingness to share your findings. Here, for example, is how the *Washington Post* began one of its telephone surveys:

> Hello, I'm (NAME), calling for the *Washington Post* public opinion poll. We're not selling anything, just doing an opinion poll on interesting subjects for the news.[41]

This introduction is short and businesslike but friendly. The interviewers have no doubt rehearsed the message countless times so that they can repeat it with confidence and professionalism.

In general, interviewers are expected to motivate the respondents. Generally it has been thought that warm, friendly interviewers who develop a good rapport with

[41] *Washington Post* Virginia Governor Poll #2, October [computer file], ICPSR04522-v1 (Horsham, Penn.: Taylor Nelson Sofres Intersearch [producer], 2005; Ann Arbor, Mich.: Inter-university Consortium for Political and Social Research [distributor], March 9, 2007), retrieved March 31, 2007, from http://www.icpsr.umich.edu/

respondents motivate them to give quality answers and to complete the survey. Yet some research has questioned the importance of rapport.[42] Friendly, neutral, "rapport-style" interviews in which interviewers give only positive feedback no matter what the response may not be good enough, especially if the questions involve difficult reporting tasks. Both types of feedback—positive ("yes, that's the kind of information we want") and negative ("that's only two things")—may improve response quality. Interviewers also may need to instruct respondents about how to provide complete and accurate information. This more businesslike, task-oriented style has been found to lead to better reporting than rapport-style interviewing.[43]

Interviewer style appears to make less difference in telephone interviews, perhaps because of the lack of visual cues the respondent can use to judge the interviewer's sincerity.[44] Even something as simple as intonation, however, may affect data quality. Interviewers whose voices go up rather than down at the end of a question appear to motivate a respondent's interest in reporting.[45]

Despite the advantages of using interviewers to improve response quality, the interviewer-respondent interaction may also bias a respondent's answers. The interviewer may give a respondent the impression that certain answers are expected or are correct. The age, gender, or race of the interviewer may affect the respondent's willingness to give honest answers. For example, on questions about race, respondents interviewed by a member of another race have been found to be more deferential to the interviewer (that is, try harder not to cause offense) than those interviewed by a member of their own race.[46] Education also has an impact on race-of-interviewer effects: less-educated blacks are more deferential than better-educated blacks, and better-educated whites are more deferential than less-educated whites.[47]

[42] See Willis J. Goudy and Harry R. Potter, "Interview Rapport: Demise of a Concept," *Public Opinion Quarterly* 39, no. 4 (1975): 529–43; and Charles F. Cannell, Peter V. Miller, and Lois Oksenberg, "Research on Interviewing Techniques," in *Sociological Methodology 1981*, ed. Samuel Leinhardt (San Francisco: Jossey-Bass, 1981), 389–437.

[43] Rogers, "Interviews by Telephone and in Person."

[44] Ibid.; and Peter V. Miller and Charles F. Cannell, "A Study of Experimental Techniques for Telephone Interviewing," *Public Opinion Quarterly* 46, no. 2 (1982): 250–69.

[45] Arpåd Barath and Charles F. Cannell, "Effect of Interviewer's Voice Intonation," *Public Opinion Quarterly* 40, no. 3 (1976): 370–73.

[46] Patrick R. Cotter, Jeffrey Cohen, and Philip B. Coulter, "Race-of-Interviewer Effects in Telephone Interviews," *Public Opinion Quarterly* 46, no. 2 (1982): 278–84; and Bruce A. Campbell, "Race of Interviewer Effects among Southern Adolescents," *Public Opinion Quarterly* 45, no. 2 (1981): 231–44.

[47] Shirley Hatchett and Howard Schuman, "White Respondents and Race-of-Interviewer Effects," *Public Opinion Quarterly* 39, no. 4 (1975–76): 523–28; and Michael F. Weeks and R. Paul Moore, "Ethnicity of Interviewer Effects on Ethnic Respondents," *Public Opinion Quarterly* 45, no. 2 (1981): 245–49.

INTERVIEWER CHARACTERISTICS. These may have a larger effect on telephone surveys than in-person surveys.[48] Because of its efficiency and because telephone interviewers, even for national surveys, do not need to be geographically dispersed, telephone interviewing requires fewer interviewers than does personal interviewing to complete the same number of interviews. Centralization of telephone interviewing operations, however, allows closer supervision and monitoring of interviewers, making it easier to identify and control interviewer problems. For both personal and telephone interviewers, training and practice is an essential part of the research process. Internet or mail surveys can avoid this bias, as both can be executed without interviewers.

PROBING. As just noted, politics is not on the top of everyone's mind. Consequently, it is often necessary to tease out responses. An interviewer can probe for additional information or clarification. He or she can gently encourage the respondent to think a bit or add more information rather than just provide an off-the-cuff answer. For example, suppose you want to know how people feel about presidential candidates. You could, as many polls do, list a number of qualities or characteristics that respondents can apply to the choices. But this technique assumes that you know what people are thinking. By contrast, if you simply ask a subject, "What do you think about Candidate X?" the first reply is often something like "Hmm . . . not much," or "She's a jerk." This may or may not be a true feeling, and in all likelihood it is not complete. Often, however, people pause a moment before responding. A trained interviewer waits a short while for the person to gather his or her thoughts. If the answer is not totally clear, the interviewer can ask for clarification. This is how the American National Election Studies, a series of major academic surveys, handles the problem. The lead-in begins, "Now I'd like to ask you about the good and bad points of the major candidates for President. . . . Is there anything in particular about [name of candidate] that might make you want to vote for him?" The questionnaire then reads:

> IF R [the respondent] SAYS THERE IS SOMETHING THAT WOULD MAKE R VOTE [for the candidate]:
>
> QUESTION:
>
> (What is that?)
>
> INTERVIEWER INSTRUCTION:
>
> {PROBE: ANYTHING ELSE? UNTIL R SAYS NO}[49]

[48] See Eleanor Singer, Martin R. Frankel, and Marc B. Glassman, "The Effect of Interviewer Characteristics and Expectations on Response," *Public Opinion Quarterly* 47, no. 1 (1983); Groves and Kahn, *Surveys by Telephone;* Dillman, *Mail and Telephone Surveys;*.and John Freeman and Edgar W. Butler, "Some Sources of Interviewer Variance in Surveys," *Public Opinion Quarterly* 40, no. 1 (1976): 79–91.

[49] Adapted from American National Election Study (ANES) 2004, "HTML Codebook Produced July 14, 2006." Accessed March 10, 2007. Available at http://sda.berkeley.edu/

Survey Type and Response Quality

The ability to obtain quality responses differs according to the type of survey used. Although with mail and drop-off surveys, an interviewer cannot probe for additional information or clarification, these surveys may have an advantage in obtaining truthful answers to threatening or embarrassing questions because anonymity can be assured and answers given in private. A mail survey also gives the respondent enough time to finish when it is convenient; this enables the respondent to check records to provide accurate information, something that is harder to arrange in telephone and personal interviews.

Disadvantages of the mail survey include problems with open-ended questions. (As we see in the next section, an open-ended question asks for the respondent's own words, as in "Is there anything in particular that you like about the Republican Party?" The respondent can say whatever he or she thinks.) Some respondents may lack writing skills or find answering at length a burden. No interviewer is present to probe for more information, clarify complex or confusing questions, motivate the respondent to answer tedious or boring questions, or control who else may contribute to or influence answers.

Personal and telephone interviews share many advantages and disadvantages with respect to obtaining quality responses, although some important differences exist. Several of the advantages of personal and telephone interviews over mail surveys stem from the presence of an interviewer. As noted earlier, an interviewer may obtain better-quality data by explaining questions, probing for more information to open-ended questions, and making observations about the respondent and his or her environment. For example, in a personal interview, the quality of furnishings and housing may be an indicator of income, and in a telephone interview, the amount of background noise might affect the respondent's concentration. In a personal interview, the interviewer may note that another household member is influencing a respondent's answers and take steps to curtail this influence. Influence by others is generally not a problem with telephone interviews, since only the respondent hears the questions. One response-quality problem that does occur with telephone interviews is that the respondent may not be giving the interviewer his or her undivided attention. This may be difficult for the interviewer to detect and correct.

Numerous studies have compared the response quality of personal and telephone interviews. One expected difference is in answers to open-ended questions. Telephone interviewers lack visual cues for probing. Thus, telephone interviews tend to be quick paced; pausing to see if the respondent adds more to an answer is more awkward on the telephone than in person. Research findings, however, have been mixed. One study found that shorter answers were given to open-ended questions in telephone interviews, especially among respondents who typically give complete and detailed responses; another study found no difference between personal and telephone interviews in the number of responses to open-ended questions.[50] Asking an open-ended question early in a telephone survey helps to relax the respondent, reduce the pace of the interview, and ensure that the respondent is thinking about his or her answers.[51]

Response quality may be lower for telephone interviews than for face-to-face interviews because of the difficulty of asking complex questions or questions with many response categories over the phone. Research has found more acquiescence, evasiveness, and extremeness in telephone survey responses than in personal survey responses. In addition, phone respondents give more contradictory answers to checklist items and are less likely to admit to problems.[52] This finding contradicts the expectation that telephone interviews result in more accurate answers to sensitive questions because of reduced personal contact.

Researchers using personal and telephone interviews have developed techniques to obtain more accurate data on sensitive topics.[53] Problems often can be avoided simply by careful wording choice. For example, for questions about socially desirable behavior, a casual approach reduces the threat by lessening the perceived importance of the topic. The question, "For whom did you vote in the last election?" could inadvertently stigmatize nonvoting. Here, once more, is how the American National Election Studies put the question:

> In talking to people about elections, we often find that a lot of people were not able to vote because they weren't registered, they were sick, or they just didn't have time. How about you—did you vote in the elections this November?[54]

In other words, giving respondents reasons for not doing something perceived as socially desirable reduces threat and may cut down on overreports of the behavior.

[50] See Groves and Kahn, *Surveys by Telephone;* and Lawrence A. Jordan, Alfred C. Marcus, and Leo G. Reeder, "Response Styles in Telephone and Household Interviewing," *Public Opinion Quarterly* 44, no. 2 (1980): 210–22.

[51] Dillman, *Mail and Telephone Surveys.*

[52] Jordan, Marcus, and Reeder, "Response Styles"; Groves and Kahn, *Surveys by Telephone.* See also Rogers, "Interviews by Telephone and in Person."

[53] For example, see Seymour Sudman and Norman M. Bradburn, *Asking Questions: A Practical Guide to Questionnaire Design* (San Francisco: Jossey-Bass, 1982), 55–86; Jerald G. Bachman and Patrick M. O'Malley, "When Four Months Equal a Year: Inconsistencies in Student Reports of Drug Use," *Public Opinion Quarterly* 45, no. 4 (1981): 536–48.

[54] American National Election Study (ANES) 2004, "HTML Codebook Produced July 14, 2006." Accessed March 10, 2007. Available at http://sda.berkeley.edu/

Interlude: The Health Care Debate and Why Question Wording Matters

Before proceeding with an examination of the dos and don'ts of question wording, we present a substantive example of the importance of question wording.

The chapter began by mentioning how groups use polls to advance their causes. The reason is obvious: public support is a potent resource in any democratic conflict. An organization that asserts the people support its objectives may have an advantage over its opponents. After all, what slogan beats "The American people want it!"? And it is usually a snap to find a poll that, if looked at in the right way, can be interpreted as proving there is a mandate for nearly any cause. The health care debate that roiled American politics during President Obama's first term is a case in point.

We pointed out in the beginning of the chapter that the Patient Protection and Affordable Care Act (aka "Obamacare") has been opposed on many grounds. Among the strongest objections is the assertion that the "people don't want it." Republicans have been especially vocal in making this point. Here, for example, is Rep. Steve King, Republican of Iowa:

> *On March 23, 2010, President Obama defied the wishes of the American people and signed ObamaCare into law. At the time, supporters of the unconstitutional law claimed that Americans would warm up to ObamaCare as they began to realize what was in it. This claim was unlikely then, and recent polling indicates that it is completely without merit now. . . . The lesson derived from these results? Americans increasingly favor repeal, and their position on this matter is hardening.*[55]

Most Democrats disagree. Although they don't explicitly argue that public opinion backs them, they predict voters will "come around" when they start benefiting from PPACA's many provisions.

Since both sides resort to polling to bolster their arguments, we might think this would be an easy dispute to settle. Alas, things are not quite so simple. Consider the results taken from several organizations in the first months of 2011. (*Note:* These questions come from polls conducted by various organizations. All entail national samples of about 1,000 individuals and were conducted during roughly the same period, in early 2011.)

- Do Americans approve of the PPACA? Not really, according to many polls taken in early 2011, one year after the reform's passage.

[55] Steve King, "Public Wants ObamaCare Repealed, and I Can Do It," *Florida Political Press* (blog), March 26, 2011. Available at http://www.floridapoliticalpress.com/2011/03/26/public-wants-obamacare-repealed-and-i-can-do-it/

Survey	Good thing/ Favor %	Bad thing/ Oppose %	Unsure %
Gallup Poll*	46	44	10
CNN/ORC§	37	59	5
Kaiser Family Foundation†	42	46	13
CBS News Poll‡	33	51	16
Average	39.5%	50%	11%

*Gallup Poll, March 18–19, 2011. *N* = 1,038 adults nationwide.

§CNN/Opinion Research Corporation Poll, March 11–13, 2011. *N* = 1,023 adults nationwide.

†Kaiser Family Foundation Health Tracking Poll, March 8–13, 2011. *N* = 1,202 adults nationwide.

‡CBS News Poll, February 11–14, 2011. *N* = 1,031 adults nationwide. "Strongly approve" and "somewhat approve" responses are combined as are disapprovals.

- Do they therefore want it repealed? Senator Jim DeMint (R-SC) thinks so:

  > *Republicans are standing with the American people who are demanding we repeal this government takeover of health care.*[56]

- But from aggregate numbers it is not clear. Overall, in fact, it appears the country is closely divided. Maybe, then, we need to dig under the surface. As the data imply, but certainly do not demonstrate, there doesn't seem to be a stampede for repeal.

Survey	Repeal %	Let it stand %	Unsure %
CBS News/*New York Times* Poll*	40	48	12
NBC News/*Wall Street Journal*§	45	46	9
CNN/Opinion Research Corporation Poll†	50	42	8
Average	45%	44.7	9.7%

*CBS News/*New York Times* Poll, January 15–19, 2011. *N* = 1,036 adults nationwide.

§NBC News/*Wall Street Journal* Poll conducted by the polling organizations of Peter Hart (D) and Bill McInturff (R), January 13–17, 2011. *N* = 1,000 adults nationwide. Those "strongly" and "not so strongly" combined; the same for those "opposed."

†CNN/Opinion Research Corporation Poll, January 14–16, 2011. *N* = 1,014 adults nationwide.

(Continued)

[56] Quoted in Daniel Sayani, "Senate Republicans Seek to Repeal ObamaCare," *New American*, February 2, 2011. Available at http://thenewamerican.com/usnews/health-care/6150-senate-republicans-seek-to-repeal-obamacare/

(Continued)

- Of course, any public official will probably respond to those he or she hears most frequently and loudly. Being a Republican from a conservative state, Senator DeMint might naturally infer a groundswell of opposition to the law. So when trying to fathom the public's mind, it helps to break down further who is saying what.

Party identification	Repeal %	Let it stand %	Unsure %
Republican	73	16	11
Independent	38	45	17
Democrat	16	77	6
Total sample	**40**	**48**	**12**

"Do you think Congress should try to repeal the health care law that was passed last year, or should they let it stand?"

Source: CBS News/*New York Times* Poll, January 15–19, 2011. *N* = 1,036 adults nationwide.

- The figures suggest that in the aggregate, the people are divided; attitudes are strongly correlated with partisanship. In order to know where the public stands, one has to take account of the various groups. Independents split slightly in favor of letting the reforms alone, and they—along with the Democrats, who are overwhelmingly in favor—counterbalance Republicans, who are more or less united in opposition. The ambiguity of public opinion on the matters deepens when examining attitudes and beliefs in more detail. Rather than simply counting "for" and "against" respondents, we should look for clues to *how* respondents interpret key terms. It's tempting, for example, to assume the response "Oppose" means "against the health care bill passed in 2010." But consider these results.

	Pro-Health Care Reform Responses			
	Favor %	Oppose—Not liberal enough %	Oppose—Too liberal %	Unsure %
CNN/Opinion Research Corporation Poll	37	13	43	7

"As you may know, a bill that makes major changes to the country's health care system became law last year. Based on what you have read or heard about that legislation, do you generally favor or generally oppose it?" If oppose: "Do you oppose that legislation because you think its approach toward health care is too liberal, or because you think it is not liberal enough?"

Source: CNN/Opinion Research Corporation Poll, March 11–13, 2011. *N* = 1,023

- Respondents were first asked whether or not they oppose the reform package, and then those who were opposed were asked if the bill was too liberal or not liberal enough. Most (43%) objected because it was too "liberal," an understandable reaction given the legislation's liberal origins and support. But note also that 13 percent understood *oppose* differently; for them the act was objectionable because it did not go far enough in the direction favored by Democrats and progressives who advocate for even greater involvement in health care than the bill provides. So again, the answer to the question, "What do the voters want?" is not entirely obvious. (Needless to say, this conclusion rests on the assumption that those surveyed understood *liberal* in the same way. That may be doubtful.)
- Survey analysts often want to gauge the level of knowledge that undergirds opinions. In early 2011, House and Senate Republicans made strenuous, widely publicized attempts to repeal PPACA, but as of August of that year they had not succeeded. Many voters, however, seemed confused. A Kaiser Foundation poll posed this question: "As far as you know, which comes closest to describing the current status of the health reform law that was passed last year? It is still the law of the land. OR, It has been repealed and is no longer law." The responses:[57]
- Still the law: 52 percent
- Has been repealed: 22 percent
- Unsure: 26 percent

Barely more than half the sample knew the law was still in force. Nearly one in five thought it had been repealed, and over a quarter were not certain. What to make of this? One implication is that some opinions may not be set in stone; if people "knew" more, their thoughts about the policy might change.

- This and the previous finding point to a potential Achilles' heel in survey research: the dependence on language. Pollsters must constantly strive to establish shared frames of reference so interviewer and subject comprehend and apply words the same way. Equally important is trying to grasp the set of beliefs (thoughts about what is "real") and evaluations of beliefs (that is, attitudes toward what is believed—"Is this a good or bad state of affairs?") that underlie spoken opinions, which are frequently uttered on the spur of the moment. Below are attitudes toward various provisions of the new law. Once these are taken into account, one has a far more nuanced understanding of popular reactions to the legislation.

(Continued)

[57] Kaiser Family Foundation, February 3–6, 2011, reported in *Polling Report: Health Policy.* Accessed March 30, 2011. Available at http://www.pollingreport.com/health.htm

(Continued)

	Favor %	Oppose %	Neither/unsure/ refused %
Employer mandate*	59	32	10
Limits on coverage§	59	34	7
Prior conditions†	50	34	17
Mandate to obtain insurance‡	31	59	10

* "Do you favor, oppose, or neither favor nor oppose a law requiring most medium-size and large companies to offer health insurance to their employees or pay money to the government as a penalty if they don't?"

§ "Do you favor, oppose, or neither favor nor oppose a law saying that an insurance company cannot stop selling health insurance to one of their customers if that person gets a serious illness?"

† "Do you favor, oppose, or neither favor nor oppose a law requiring insurance companies to sell health insurance to a person who is currently sick or has had a serious illness in the past, which would probably cause most Americans to pay more for health insurance?"

‡ "Do you favor, oppose, or neither favor nor oppose a law that would require every American to have health insurance, or pay money to the government as a penalty if they do not, unless the person is very poor?"

Source: AP-GfK Poll conducted by GfK Roper Public Affairs & Corporate Communications, January 5–10, 2011. *N* = 1,001 adults nationwide.

- Except for the personal mandate, some provisions of the reform seem acceptable to perhaps a majority or plurality of Americans. The exception, the requirement that everyone purchase some form of health insurance, is wildly unpopular, perhaps because it touches on issues of individual freedom and responsibility. Whatever the case, it seems to be a stretch to argue that voters oppose the reforms in their entirety.
- It is also important to inquire into the policy implications of public opinion. One proposal for getting rid of Obamacare is to "defund" it by choking off appropriations necessary for the law's implementation. But many citizens, even Republicans, seem to be wary of this approach.

Party identification	Approve %	Disapprove %	Unsure %
Republican	57	34	10
Independent	38	49	13
Democrat	12	82	6
Average	35	55	10

Question: "Some members of Congress have said they may stop funding for the new health care law. Regardless of how you feel about the new health care legislation, would you approve or disapprove if Congress stopped funding for the new health care law?"

Source: CBS News Poll, February 11–14, 2011. *N* = 1,031 adults nationwide.

- Finally, since all of these data illustrate the role language plays in decoding poll reports, we conclude with an issue that comes up later: "leading" questions, or questions phrased in such a manner as to intentionally bias or encourage a response in one direction or another. A Fox News survey asked its participants this question:

> *Some Americans choose not to buy health insurance even though they can afford it. The president's plan requires all Americans who can afford it to have some form of health insurance or else pay a penalty. Failure to pay the penalty would result in an even larger fine, a jail sentence of up to one year, or both. Do you think the government should be able to require all Americans who can afford it to have health insurance or pay a penalty, or not?*[58]

The results were as follows:

- "Yes," government should be able to require participation: 28 percent
- "No," should not: 69 percent
- "Unsure": 4 percent

The upshot, then, is that how one frames an issue partly determines the public's expressed or verbalized stances on it. Advocacy groups, of course, have to take into account this phenomenon, but as social scientists we have to be aware of it as well. The following sections describe in more detail some considerations when writing or evaluating questionnaires.

Question Wording

The central problem of survey and interview research is that the procedures involve a structured interaction between a social scientist and a subject. This is true even if the method used involves indirect contact, such as a mail or an Internet survey. Since the whole point of survey research is to accurately measure people's attitudes, beliefs, and behavior by asking them questions, we need to spend time discussing good and bad questions. Good questions prompt accurate answers; bad questions provide inappropriate stimuli and result in unreliable or inaccurate responses. When writing questions, researchers should use objective and clear wording. Failure to do so may result in incomplete questionnaires and meaningless data for the researcher. The basic rule is this: *the target subjects must be able to understand and in principle have access to the requested information.* Try to put yourself in the respondent's place. Would an ordinary citizen, for example, be able to reply meaningfully to the question, "What's your opinion of the recently passed amendments to the Import/Export Bank Authorization?"

[58] FOX News/Opinion Dynamics Poll, December 14–15, 2010, *Polling Report: Health Policy.* Accessed March 20, 2011. Available at http://www.pollingreport.com/health.htm. N = 900 registered voters nationwide.

OBJECTIVITY AND CLARITY. Certain types of questions make it difficult for respondents to provide reliable, accurate responses. These include double-barreled, ambiguous, and leading questions. A **double-barreled question** is really two questions in one, such as "Do you agree with the statement that the situation in Iraq is deteriorating and that the United States should increase the number of troops in Iraq?" How does a person who believes that the situation in Iraq is deteriorating but who does not wish an increase in troops answer this question? Or someone who doesn't feel the situation is worse but nevertheless believes that more troops would be advisable? And how does the researcher interpret an answer to such a question? It is not clear whether the respondent meant for his or her answer to apply to both components or whether one component was given precedence over the other.

Despite a conscious effort by researchers to define and clarify concepts, words with multiple meanings or interpretations may creep into questions. An ambiguous question is one that contains a concept that is not defined clearly. An example would be the question, "What is your income?" Is the question asking for family income or just the personal income of the respondent? Is the question asking for earned income (salary or wages), or should interest and stock dividends be included? Ambiguity also may result from using the word *he*. Are respondents to assume that *he* is being used generically to refer both to men and women or to men only? If a respondent interprets the question as applying only to men and would respond differently for women, it would be a mistake for the researcher to conclude that the response applies to all people.[59]

Researchers must avoid asking leading questions. A **leading question**, sometimes called a reactive question, encourages respondents to choose a particular response because the question indicates that the researcher expects it. The question, "Don't you think that global warming is a serious environmental problem?" implies that to think otherwise would be unusual. Word choice may also lead respondents. Research has shown that people are more willing to help "the needy" than those "on welfare." Asking people if they favor "socialized medicine" rather than "national health insurance" is bound to decrease affirmative responses. Moreover, linking personalities or institutions to issues can affect responses. For example, whether or not a person liked the governor would affect responses to the following question: "Would you say that Governor Burnett's program for promoting economic development has been very effective, fairly effective, not too effective, or not effective at all?"[60] Polls conducted by political organizations and politicians often include

[59] Margrit Eichler, *Nonsexist Research Methods: A Practical Guide* (Winchester, Mass.: Allen and Unwin, 1988), 51–52.

[60] Charles H. Backstrom and Gerald Hursh-Cesar, *Survey Research*, 2nd ed. (New York: Wiley, 1981), 142, 146.

leading questions. For example, a 1980 poll for the Republican National Committee asked, "Recently the Soviet armed forces openly invaded the independent country of Afghanistan. Do you think the U.S. should supply military equipment to the rebel freedom fighters?"[61] Before accepting any interpretation of survey responses, we should check the full text of a question to make sure that it is neither leading nor biased.

Indeed, some campaigns, parties, and political organizations have begun converting survey research into a form of telemarketing through a technique called a **push poll**. Interviewers, supposedly representing a research organization, feed respondents (often) false and damaging information about a candidate or cause under the guise of asking a question. The caller may ask, for example, "Do you agree or disagree with Candidate X's willingness to tolerate terrorism in our country?" The goal, of course, is not to conduct research but to use innuendo to spread rumors and lies.

Questions should be stated in such a way that they can produce a variety of responses. If you simply ask, "Do you favor cleaning up the environment—yes or no?" almost all the responses will surely be yes. At the same time, the alternatives themselves should encourage thoughtful replies. For instance, if the responses to the question, "How would you rate President Obama's performance so far?" are (1) great, (2) somewhere between great and terrible, and (3) terrible, you probably are not going to find very much variation, since the list practically demands that respondents pick choice (2). Also, an alternative should be available for each possible situation. For example, response options for the question, "For whom did you vote in the 2008 presidential election?" should list John McCain and Barack Obama, as well as other candidates (for example, Ralph Nader) and certainly should include generic "other" and "did not vote" options. (The section on "question type" discusses this topic in more depth.)

Use of technical words, slang, and unusual vocabulary should be avoided, since respondents may misinterpret their meaning. Questions including words with several meanings will result in ambiguous answers. For example, the answer to the question, "How much bread do you have?" depends on whether the respondent thinks of bread as coming in loaves or dollar bills. The use of appropriate wording is especially important in cross-cultural research. For researchers to compare answers across cultures, questions should be equivalent in meaning. For example, the question, "Are you interested in politics?" may be interpreted as "Do you vote in elections?" or "Do you belong to a political party?" The interpretation would depend on the country or culture of the respondent.

Attention to these basic guidelines for question wording increases the probability that respondents will interpret a question consistently and as intended, yielding

[61] Republican National Committee, *1980 Official Republican Poll on U.S. Defense and Foreign Policy.*

reliable and valid responses. Luckily, every researcher does not have to formulate questions anew. We discuss archival sources of survey questions later in this chapter.

Question Type

The form or type of question as well as its specific wording is important. There are two basic types of questions: closed-ended and open-ended. A **closed-ended question** provides respondents with a list of responses from which to choose. "Do you agree or disagree with the statement that the government ought to do more to help farmers?" and "Do you think that penalties for drunk driving are too severe, too lenient, or just about right?" are examples of closed-ended questions.

A variation of the closed-ended question is a question with multiple choices for the respondent to accept or reject. A question with multiple choices is really a series of closed-ended questions. Consider the following example: "Numerous strategies have been proposed concerning the federal budget deficit. Please indicate whether you find the following alternatives acceptable or unacceptable: (a) raise income taxes, (b) adopt a national sales tax, (c) reduce military spending, (d) reduce spending on domestic programs."

Nominal-level measures ought to consist of categories that are exhaustive and mutually exclusive; that is, the categories should include all the possibilities for the measure, and every respondent should fit in one and only one category. Researchers use "other" when they are unable to specify all alternatives, or when they expect very few of their observations to fall into the "other" category but want to provide an option for respondents who do not fall into one of the labeled categories (respondents may fail to complete surveys with questions that don't apply to them). For example, questions asking for a person's religion often include "other" as an option. If using "other" as a category, you should check your data to make sure that only a relatively few observations fall into it. Otherwise, subsequent data analysis will not be very meaningful.

In an **open-ended question**, the respondent is not provided with any answers from which to choose. The respondent or interviewer writes down the answer. An example of an open-ended question is, "Is there anything in particular about BARACK OBAMA that might make you want to vote for him?"[62]

CLOSED-ENDED QUESTIONS: ADVANTAGES AND DISADVANTAGES. The main advantage of a closed-ended question is that it is easy

[62] American National Election Study (ANES) 2008, "HTML Codebook Produced April 12, 2011." Available at http://sda.berkeley.edu/D3/NES08new/Doc/hcbk.htm

to answer and takes little time. Also, the answers can be precoded (that is, assigned a number) and the code then easily transferred from the questionnaire to a computer. Another advantage is that answers are easy to compare, since all responses fall into a fixed number of predetermined categories. These advantages aid in the quick statistical analysis of data. With open-ended questions, by contrast, the researcher must read each answer, decide which answers are equivalent, decide how many categories or different types of answers to code, and assign codes before the data can be computerized.

Another advantage of closed-ended questions over open-ended ones is that respondents are usually willing to respond on personal or sensitive topics (for example, income, age, frequency of sexual activity, or political views) by choosing a category rather than stating the actual answer. This is especially true if the answer categories include ranges. Finally, closed-ended questions may help clarify the question for the respondent, thus avoiding misinterpretations of the question and unusable answers for the researcher.

Critics of closed-ended questions charge that they force a respondent to choose an answer category that may not accurately represent his or her position. Therefore, the response has less meaning and is less useful to the researcher. Also, closed-ended questions often are phrased so that a respondent must choose between two alternatives or state which one is preferred. This may result in an oversimplified and distorted picture of public opinion. A closed-ended question allowing respondents to pick more than one response (for example, with instructions to choose all responses that apply) may be more appropriate in some situations. The information produced by such a question indicates which choices are acceptable to a majority of respondents. In fashioning a policy that is acceptable to most people, policy makers may find this knowledge much more useful than simply knowing which alternative a respondent prefers.

Just as the wording of a question may influence responses, so too may the wording of response choices. Changes in the wording of question responses can result in different response distributions. Two questions from the 1960s concerning troop withdrawal from Vietnam illustrate this problem.[63] A June 1969 Gallup Poll question asked,

> President Nixon has ordered the withdrawal of 25,000 United States troops from Vietnam in the next three months. How do you feel about this—do you think troops should be withdrawn at a faster rate or a slower rate?

[63] John P. Dean and William Foote Whyte, "How Do You Know If the Informant Is Telling the Truth?" in *Elite and Specialized Interviewing*, ed. Lewis Anthony Dexter (Evanston, Ill.: Northwestern University Press, 1970), 127.

The answer "same as now" was not presented but was accepted if given. The response distribution was as follows: faster, 42 percent; same as now, 29 percent; slower, 16 percent; no opinion, 13 percent.

Compare the responses with those to a September–October 1969 Harris Poll in which respondents were asked,

> In general, do you feel the pace at which the president is withdrawing troops is too fast, too slow, or about right?

Responses to this question were as follows: too slow, 28 percent; about right, 49 percent; too fast, 6 percent; no opinion, 18 percent.

Thus, support for presidential plans varied from 29 to 49 percent. The response depended on whether respondents were directly given the choice of agreeing with presidential policy or had to mention such a response spontaneously.

Response categories may also contain leading or biased language and may not provide respondents with equal opportunities to agree or disagree. Response distributions may be affected by whether the researcher asks a **single-sided question**, in which the respondent is asked to agree or disagree with a single substantive statement, or a **two-sided question**, which offers the respondent two substantive choices. An example of a one-sided question is

> Do you agree or disagree with the idea that the government should see to it that every person has a job and a good standard of living?

An example of a two-sided question is

> Do you think that the government should see to it that every person has a job and a good standard of living, or should it let each person get ahead on his or her own?

With a single-sided question, a larger percentage of respondents tend to agree with the statement given. Forty-four percent of the respondents to the single-sided question given above agreed that the government should guarantee employment, whereas only 30.3 percent of the respondents to the two-sided question chose this position.[64] Presenting two substantive choices has been found to reduce the proportion of respondents who give no opinion.[65]

Closed-ended questions may provide inappropriate choices, thus leading many respondents to not answer or to choose the "other" category. Unless space is

[64] Raymond L. Gordon, *Interviewing: Strategy, Techniques, and Tactics* (Homewood, Ill.: Dorsey, 1969), 18.

[65] Dexter, *Elitê and Specialized Interviewing*, 17.

provided to explain "other" (which then makes the question resemble an open-ended one), it is anybody's guess what "other" means. Another problem is that errors may enter into the data if the wrong response code is marked. With no written answer, inadvertent errors cannot be checked. A problem also arises with questions having a great many possible answers. It is time-consuming to have an interviewer read a long list of fixed responses that the respondent may forget. A solution to this problem is to use a response card. Responses are typed on a card that is handed to the respondent to read and choose from.

OPEN-ENDED QUESTIONS: ADVANTAGES AND DISADVANTAGES. Unstructured, free-response questions allow respondents to state what they know and think. They are not forced to choose between fixed responses that do not apply. Open-ended questions allow respondents to tell the researcher how they define a complex issue or concept. As one survey researcher in favor of open-ended questions pointed out,

> Presumably, although this is often forgotten, the main purpose of an interview, the most important goal of the entire survey profession, is to let the respondent have his say, to let him tell the researcher what he means, not vice versa. If we do not let the respondent have his say, why bother to interview him at all?[66]

Sometimes researchers are unable to specify in advance the likely responses to a question. In this situation, an open-ended question is appropriate. Open-ended questions are also appropriate if the researcher is trying to test the knowledge of respondents. For example, respondents are better able to *recognize* names of candidates in a closed-ended question (that is, pick the candidates from a list of names), than they are able to *recall* names in response to an open-ended question about candidates. Using only one question or the other would yield an incomplete picture of citizens' awareness of candidates.

Paradoxically, a disadvantage of the open-ended question is that respondents may respond too much or too little. Some may reply at great length about an issue—a time-consuming and costly problem for the researcher. On the other hand, if open-ended questions are included on mail surveys, some respondents with poor writing skills may not answer, which may bias responses. Thus, the use of open-ended questions depends on the type of survey. Another problem is that interviewers may err in recording a respondent's answer. Recording answers verbatim is tedious. Furthermore, unstructured answers may be difficult to code, interpretations of answers may vary (affecting the reliability of data), and processing answers may become

[66] Patricia J. Labaw, *Advanced Questionnaire Design* (Cambridge, Mass.: Abt Books, 1980), 132.

time-consuming and costly. For these reasons, open-ended questions are often avoided—although unnecessarily, in Patricia Labaw's opinion:

> I believe that coding costs have now been transferred into data-processing costs. To substitute for open questions, researchers lengthen their questionnaires with endless lists of multiple choice and agree/disagree statements, which are then handled by sophisticated data-processing analytical techniques to try to massage some pattern or meaning out of the huge mass of pre-coded and punched data. I have found that a well-written open-ended question can eliminate the need for several closed questions, and that subsequent data analysis becomes clear and easy compared to the obfuscation provided by data massaging.[67]

Question Order

The order in which questions are presented to respondents may also influence the reliability and validity of answers. Researchers call this the **question-order effect.** In ordering questions, the researcher should consider the effect on the respondent of the previous question, the likelihood of the respondent's completing the questionnaire, and the need to select groups of respondents for certain questions. In many ways, answering a survey is a learning situation, and previous questions can be expected to influence subsequent answers. This presents problems as well as opportunities for the researcher.

The first several questions in a survey are usually designed to break the ice. They are general questions that are easy to answer. Complex, specific questions may cause respondents to terminate an interview or not complete a questionnaire because they think it will be too hard. Questions on personal or sensitive topics usually are left to the end. Otherwise, some respondents may suspect that the purpose of the survey is to check up on them rather than to find out public attitudes and activities in general. In some cases, however, it may be important to collect demographic information first. In a study of attitudes toward abortion, one researcher used demographic information to infer the responses of those who terminated the interview. She found that older, low-income women were most likely to terminate the interview on the abortion section. Since their group matched those who completed the interviews and who were strongly opposed to abortion, she concluded that termination expressed opposition to abortion.[68]

One problem to avoid is known as a **response set**, or straight-line responding. A response set may occur when a series of questions have the same answer choices.

[67] Ibid., 132–33.

[68] Ibid., 117.

Respondents who find themselves agreeing with the first several statements may skim over subsequent statements and check "agree" on all. This is likely to happen if statements are on related topics. To avoid the response-set phenomenon, statements should be worded so that respondents may agree with the first, disagree with the second, and so on. This way the respondents are forced to read each statement carefully before responding.

Additional question-order effects include saliency, redundancy, consistency, and fatigue.[69] Saliency is the effect that specific mention of an issue in a survey may have in causing a respondent to mention the issue in connection with a later question: the earlier question brings the issue forward in the respondent's mind. For example, a researcher should not be surprised if respondents mention crime as a problem in response to a general question on problems affecting their community if the survey had earlier asked them about crime in the community. Redundancy is the reverse of saliency. Some respondents, unwilling to repeat themselves, may not say crime is a problem in response to the general query if earlier they had indicated that crime was a problem. Respondents may also strive to appear consistent. An answer to a question may be constrained by an answer given earlier. Finally, fatigue may cause respondents to give perfunctory answers to questions late in the survey. In lengthy questionnaires, response-set problems often arise due to fatigue.[70]

The "learning" that takes place during an interview may be an important aspect of the research being conducted. The researcher may intentionally use this process to find out more about the respondent's attitudes and potential behavior. Labaw referred to this as "leading" the respondent and noted it is used "to duplicate the effects of information, communication and education on the respondent in real life."[71] The extent of a respondent's approval or opposition to an issue may be clarified as the interviewer introduces new information about the issue.

In some cases, such education *must* be done to elicit needed information on public opinion. For example, one study set out to evaluate public opinion on ethical issues in biomedical research.[72] Because the public is generally uninformed about these issues, some way had to be devised to enable respondents to make meaningful judgments. The researchers developed a procedure for presenting "research vignettes." Each vignette described or illustrated a dilemma actually encountered in

[69] Norman M. Bradburn and William M. Mason, "The Effect of Question Order on Responses," *Journal of Marketing Research* 1, no. 4 (1964): 57–61.

[70] Regula Herzog and Jerald G. Bachman, "Effects of Questionnaire Length on Response Quality," *Public Opinion Quarterly* 45, no. 4 (1981): 549–59. Available at http://www.uta.edu/faculty/richarme/MARK%205338/Articles/Herzog.pdf

[71] Labaw, *Advanced Questionnaire Design,* 122.

[72] Glen D. Mellinger, Carol L. Huffine, and Mitchell B. Balter, "Assessing Comprehension in a Survey of Public Reactions to Complex Issues," *Public Opinion Quarterly* 46, no. 1 (1982): 97–109.

biomedical research. A series of questions asking respondents to make ethical judgments followed each vignette. Such a procedure was felt to provide an appropriate decision-making framework for meaningful, spontaneous answers and a standard stimulus for respondents. A majority of persons, even those with less than a high school education, were able to express meaningful and consistent opinions.

If there is no specific reason for placing questions in a particular order, researchers may vary questions randomly to control question-order bias. Computerized word processing of questionnaires makes this an easier task.[73]

Question order also becomes an important consideration when the researcher uses a **branching question**, which sorts respondents into subgroups and directs these subgroups to different parts of the questionnaire, or a **filter question**, which screens respondents from inappropriate questions. For example, a marketing survey on new car purchases may use a branching question to sort people into several groups: those who bought a car in the past year, those who are contemplating buying a car in the next year, and those who are not anticipating buying a car in the foreseeable future. For each group, a different set of questions about automobile purchasing may be appropriate. A filter question is typically used to prevent the uninformed from answering questions. For example, respondents in the 1980 National Election Study were given a list of presidential candidates and asked to mark those names they had never heard of or didn't know much about. Respondents were then asked questions only about those names that they hadn't marked.

Branching and filter questions increase the chances for interviewer and respondent error.[74] Questions to be answered by all respondents may be missed. However, careful attention to questionnaire layout, clear instructions to the interviewer and the respondent, and well-ordered questions will minimize the possibility of confusion and lost or inappropriate information.

Questionnaire Design

The term **questionnaire design** refers to the physical layout and packaging of the questionnaire. An important goal of questionnaire design is to make the questionnaire attractive and easy for the interviewer and the respondent to follow. Good design increases the likelihood that the questionnaire will be completed properly. Design may also make the transfer of data from the questionnaire to the computer easier.

[73] William D. Perrault Jr., "Controlling Order-Effect Bias," *Public Opinion Quarterly* 39, no. 4 (1975): 544–51.

[74] Donald J. Messmer and Daniel T. Seymour, "The Effects of Branching on Item Nonresponse," *Public Opinion Quarterly* 46, no. 2 (1982): 270–77.

Design considerations are most important for mail questionnaires. First, the researcher must make a favorable impression based almost entirely on the questionnaire materials mailed to the respondent. Second, because no interviewer is present to explain the questionnaire to the respondent, a mail questionnaire must be self-explanatory. Poor design increases the likelihood of response error and nonresponse. Whereas telephone and personal interviewers can and should familiarize themselves with questionnaires before administering them to a respondent, the recipient of a mail questionnaire cannot be expected to spend much time trying to figure out a poorly designed form.

Using Archived Surveys

Now that you have a better idea of what surveys are and how they are properly designed and administered, it is important to understand the costs and benefits of designing and administering your own survey to collect data versus using survey questions written by or data collected by someone else. Because of the high costs involved in designing and administering a survey and the concerns about validity and reliability of the data, most students who want to design their own survey would be remiss if they did not at least consult existing surveys. In this section, we explain how you can search for survey questions and data in archives and what you can expect to find.

Advantages of Using Archived Surveys

Although some students can rely on funding from their universities for research (usually a couple hundred dollars to defray costs), most students will not have access to such funds. Without funding, most students wishing to collect survey data to analyze in a research paper will turn to the least expensive options available. The most popular source of survey data for students is a sample of undergraduate students. These efforts generally involve less expensive collection measures like in-class group surveys or surveys conducted via e-mail or the Internet. Given the limitations on available resources, these are acceptable choices, and students designing and administering their own surveys will undoubtedly learn a great deal about the pitfalls of survey research when choosing this option. Firsthand experience can be invaluable to fully understanding survey design and administration, and the experience cannot be replicated simply by reading a textbook on the subject.

One of the biggest drawbacks of students designing and administering their own surveys is that the questions, the survey form, and the administration will likely be of quite low quality without considerable input from an instructor or an advisor. Although this is not a problem if a survey is intended to be a learning exercise, students hoping to collect high-quality data from their own survey might be disappointed with the results. To be able to make valid conclusions based on the data, students should consider instead using survey questions written by professionals. Such questions are widely available for free to students through publicly available archives. In chapter 9, we discussed the availability of preprocessed or preanalyzed data in regard to document analysis. The data to which we referred are useful to students because someone else has already worked with the raw data, or answers to survey questions, and produced results in the form of tables, figures, or statistical output. In this chapter, we focus instead on the survey questions and the raw survey data, or the unvarnished answers to survey questions. These data can be more difficult to work with, but they will lend greater flexibility to students in analyzing data and making conclusions.

There are many advantages to drawing upon professional surveys for use in your own. First and foremost, imagine that you are ready to embark on a research project for which you will need survey data. You would be safe in assuming that using data someone else collected would save a great deal of time, effort, and resources. The key, of course, is finding data that will allow you to test your hypotheses and answer your research questions. Fortunately, myriad data archives are publicly available, with surveys and sample data collected from many different populations and about many topics. Second, using a professionally designed survey should lead to better data, collected from answers to well-written questions. Having taken a data analysis course, and read this chapter, students should have a good idea of what to look for in a survey design to determine the quality of the questions and, subsequently, the data. Third, using a professional survey can help convince readers that the results reported in a research report are valid because the questions used to collect the data have been used by others, potentially in published, peer-reviewed work.

A great place to start is Cornell University's CISER Data Archive.[75] The data archive has links to some of the most well-known surveys that cover social and political content, including the American National Election Studies, the General Social Survey, the Maxwell Poll, and many others, both domestic and foreign. The linked sites offer a wealth of polling reports and data that will allow students to both explore the world of survey research and find useful questions and data for use in their own projects.

Interviewing

Interviewing is simply the act of asking individuals a series of questions and recording their responses. The interaction may be face to face or over the phone.[76] In some cases, the interviewer asks a predetermined set of questions; in others,

[75] Cornell University, CISER Data Archive. Accessed February 12, 2015. Available at http://www.ciser.cornell.edu/info/polls.shtml

[76] Occasionally, the investigator may obtain the information from some form of written communication.

the discussion may be more spontaneous or freewheeling; and in still others, both structured and unstructured formats are used. The key is that an interview, like a survey, depends on the participants sharing a common language and understanding of terms. And whereas a formal questionnaire, once constructed, limits opportunities for empathetic understanding, an in-depth interview gives the interviewer a chance to probe, to clarify, to search for deeper meanings, to explore unanticipated responses, and to assess intangibles such as mood and opinion intensity.

Perhaps one of the finest examples of the advantages of extended interviews is Robert E. Lane's study of fifteen "urban male voters."[77] Although the sample seems small, Lane provided evidence that it is representative of working- and middle-class men living in an Atlantic seaboard town he calls "Eastport." More important for his purposes, his method—a series of extended and taped individual interviews lasting a total of ten to fifteen hours per subject—allowed him to delve into the political consciousness of his subjects in a way no cut-and-dried survey could.

Among many other topics, Lane explored these men's attitudes toward "equality" and a hypothetical movement toward an equalitarian society. Of course, he could have written survey-type questions that would have asked respondents if they agreed or disagreed with this or that statement. Instead, he let his subjects speak for themselves. And what he found turned out to be very interesting and unexpected:

> The upper working class, and the lower middle class, support specific measures embraced in the formula "welfare state," which have equalitarian consequences. But, so I shall argue, many members of the working class do not want equality. They are afraid of it. In some ways they already seek to escape from it.[78]

Why did he come to this startling conclusion? Because during his long interviews he uncovered several latent patterns in the men's thinking, patterns that would have been difficult to anticipate and virtually impossible to garner from a standardized questionnaire. For example, when asked about the desirability of greater equality of opportunity and income, one man, Sullivan, a railroad firefighter, said,

> I think it's hard. . . . Supposing I came into a lot of money, and I moved into a nice neighborhood—class—maybe I wouldn't know how to act then. I think it's very hard, because people know that you just—word gets around that you . . . never had it before you got it now. Well, maybe they wouldn't like you . . . maybe you don't know how to act.[79]

[77] Robert E. Lane, "The Fear of Equality," *American Political Science Review* 53, no. 1 (1959): 35 -51. The complete results of Lane's work are found in his *Political Ideology: Why the Common Man Believes What He Does* (New York: Free Press, 1963).

[78] Lane, "The Fear of Equality," 35.

[79] Ibid., 46.

Lane termed this response a concern with "social adjustment" and found that others shared the sentiment. He discovered another source of unease: those in the lower classes would not necessarily deserve a "promotion" up the social ladder. Thus, Ruggiero, a maintenance worker, believed "There's laziness, you'll always have lazy people," while another man said,

> But then you get a lot of people who don't want to work; you got welfare. People will go on living on that welfare—they're happier than hell. Why should they work if the city will support them?[80]

The research uncovered similar fears that Lane's subjects experienced when envisioning an equalitarian society. They believed such a society would be unfair to "meritorious elites," would entail the loss of "goals" (if everyone is equal, why work?), and would cause society to "collapse."

Our quick review of Lane's research should not be interpreted as an argument that his is the definitive study. One could, in fact, interpret some of the men's statements quite differently. But the men of Eastport, like all citizens, had mixed, frequently contradictory thoughts, and only after hours of conversation and considerable analysis of the transcripts could Lane begin to classify and make sense of them.

The Ins and Outs of Interviewing

Interviewing, as we use the term, differs substantially from the highly structured, standardized format of survey research.[81] There are many reasons for this difference. First, a researcher may lack sufficient understanding of events to be able to design an effective, structured survey instrument or schedule of questions. The only way for researchers to learn about certain events is to interview participants or eyewitnesses directly. Second, a researcher is usually especially interested in an interviewee's own interpretation of events or issues and does not want to lose the valuable information that an elite "insider" may possess by unduly constraining responses. As one researcher put it, "A less structured format is relatively exploratory and stresses subject rather than researcher definitions of a problem."[82]

Finally, some people, especially elites or those in positions of high standing or power, may resent being asked to respond to a standardized set of questions. In her

[80] Ibid., 44–45.

[81] There are exceptions to this general rule, however. See John Kessel, *The Domestic Presidency* (Belmont, Calif.: Duxbury, 1975). Kessel administered a highly structured survey instrument to Richard Nixon's Domestic Council staff.

[82] Joseph A. Pika, "Interviewing Presidential Aides: A Political Scientist's Perspective," in *Studying the Presidency*, ed. George C. Edwards III and Stephen J. Wayne (Knoxville: University of Tennessee Press, 1982), 282.

study of Nobel laureates, for example, Harriet Zuckerman found that her subjects soon detected standardized questions. Because these were people used to being treated as individuals with minds of their own, they resented "being encased in the straightjacket of standardized questions."[83] Therefore, those who interview elites often vary the order in which topics are broached and the exact form of questions asked from interview to interview.

In this method, eliciting valid information may require variability in approaches.[84] Interviewing is not as simple as lining up a few interviews and chatting for a while. The researcher using the in-depth interview technique must consider numerous logistical and methodological questions. Advance preparation is extremely important. The researcher should study all available documentation of events and pertinent biographical material before interviewing a member of an elite group. Advance preparation serves many purposes. First, it saves the interviewee's time by eliminating questions that can be answered elsewhere. The researcher may, however, ask the interviewee to verify the accuracy of the information obtained from other sources. Second, it gives the researcher a basis for deciding what questions to ask and in what order. Third, advance preparation helps the researcher to interpret and understand the significance of what is being said, to recognize a remark that sheds new light on a topic, and to catch inconsistencies between the interviewee's version and other versions of events. Fourth, the researcher's serious interest in the topic impresses the interviewee. At no time, however, should the researcher dominate the conversation to show off his or her knowledge. Finally, good preparation buoys the confidence of the novice researcher who is interviewing important people.

The ground rules that will apply to what is said in an interview should be made clear at the start.[85] When the interview is requested, and at the beginning of the interview itself, the researcher should ask whether confidentiality is desired. If he or she promises confidentiality, the researcher should be careful not to reveal a person's identity in written descriptions. A touchy problem in confidentiality may arise if questions are based on previous interviews. It may be possible for an interviewee to guess the identity of the person whose comments must have prompted a particular question.

A researcher may desire and promise confidentiality in the hope that the interviewee will be more candid.[86] Interviewees may request confidentiality if they fear they may reveal something damaging to themselves or to others. Some persons

[83] Harriet Zuckerman, "Interviewing an Ultra-Elite," *Public Opinion Quarterly* 36, no. 2 (1972): 167.

[84] Gordon, *Interviewing: Strategy, Techniques, and Tactics,* 49–50.

[85] Dom Bonafede, "Interviewing Presidential Aides: A Journalist's Perspective," in *Studying the Presidency,* ed. George C. Edwards III and Stephen J. Wayne (Knoxville: University of Tennessee Press, 1982), 269.

[86] Richard F. Fenno Jr., *Home Style: House Members in Their Districts* (Boston: Little, Brown, 1978), 280.

HELPFUL HINTS

Ask the Right Questions

The importance of thoroughly researching a topic before conducting elite interviews cannot be stressed enough. In addition to the guidelines discussed in the text, ask yourself this question: Can the information be provided only (or at least most easily) by the person being interviewed? If you can obtain the answers to your questions from newspapers or books, for example, then it is pointless to take up someone's time going over what is (or should be) already known. If, however, the subject believes that only she or he can help you, then you are more likely to gain her or his cooperation. Looking and acting professional is absolutely essential. So, for example, do not arrive at the interview wearing a ball cap or without paper and pen.

Check out more Helpful Hints at **edge.sagepub.com/johnson8e**

may want to approve anything written based on what they have said. In any event, it often is beneficial to the researcher to give interviewees a chance to review what has been written about them and the opportunity to clarify and expand on their comments. Sometimes a researcher and an interviewee may disagree over the content or interpretation of the interview. If the researcher has agreed to let an interviewee have final say on the use of an interview, the agreement should be honored. Otherwise, the decision is the researcher's—to be made in light of the needs of the investigation.

Sometimes, gaining access to influential people is difficult. They may want further information about the purpose of the research or need to be convinced of the professionalism of the researcher. Furthermore, many have "gatekeepers" who limit access to their bosses. It is advisable to obtain references from people who are known to potential interviewees. For example, suppose you want to talk to a few state senators. Try getting your own representative to make a few phone calls or write an introductory letter. Sometimes a person who has already been interviewed will assist a researcher in gaining access to other elites. Having a letter of recommendation or introduction from someone who knows the subject can be extremely helpful in this regard.

Whom to interview first is largely a theoretical decision. Interviewing persons of lesser importance in an event or of lower rank in an organization first allows a researcher to become familiar with special terminology used by an elite group and more knowledgeable about a topic before interviewing key elites. It also may bolster a researcher's experience and confidence. Lower-level personnel may be more candid and revealing about events because they are able to observe major participants and have less personal involvement. Talking to superiors first, however, may indicate to subordinates that being interviewed is permissible. Moreover, interviewing key elites first may provide a researcher with important information early on and make subsequent interviewing more efficient. Other factors, such as age of respondents, availability, and convenience, may also affect interview order.

A tape recorder or handwritten notes may be used to record an interview. There are numerous factors to consider in choosing between the two methods. Tape recording allows the researcher to think about what the interviewee is saying, to check notes, and to formulate follow-up questions. If the recording is clear, it removes the possibility of error about what is said. Disadvantages include the fact that everything is recorded. The material must then be transcribed (an expense) and read before useful data are at hand. Much of what is transcribed will not be useful—a problem of elite interviewing in general. A tape recorder may make some interviewees uncomfortable, and they may not be candid even if promised confidentiality; there can be no denying what is recorded. Sometimes the researcher will be unfamiliar with recording equipment and will appear awkward.

Many researchers rely on handwritten notes taken during an interview. It is important to write up interviews in more complete form soon after the interview, while it is still fresh in the researcher's mind. Typically this takes much longer than the interview itself, so enough time should be allotted. Only a few interviews should be scheduled in one day; after two or three, the researcher may not be able to recollect individual conversations distinctly. How researchers go about conducting interviews will vary by topic, by researcher, and by respondent.

Although interviews are usually not rigidly structured, researchers still may choose to exercise control and direction in an interview. Many researchers conduct a semistructured or flexible interview—what is called a **focused interview**—when questioning elites. They prepare an interview guide, including topics, questions, and the order in which they should be raised. Sometimes alternative forms of questions may be prepared. Generally the more exploratory the purpose of the research, the less topic control exercised by the researcher. Researchers who desire information about specific topics should communicate this need to the person being interviewed and exercise enough control over the interview to keep it on track.

Establishing the meaningfulness and validity of the interview data is important. The data may be biased by the questions and actions of the interviewer. Interviewees may give evasive or untruthful answers. As noted earlier, advance preparation may help an interviewer recognize remarks that differ from established fact.

Examining the remarks' plausibility, checking for internal consistency, and corroborating them with other interviewees also may determine the validity of an interviewee's statements. John P. Dean and William Foote Whyte argued that a researcher should understand an interviewee's mental set and how it might affect his or her perception and interpretation of events.[87] Raymond L. Gordon stressed the value of being able to empathize with interviewees to understand the meaning of what they are saying.[88] Lewis Dexter warned that interviews should be conducted only if "interviewers have enough relevant background to be sure that they can make sense out of interview conversations or . . . there is reasonable hope of being able to . . . learn what is meaningful and significant to ask."[89]

Despite the difficulties, interviewing is an excellent form of data collection, particularly in exploratory studies or when thoughts and behaviors can be described or expressed only by those who are deeply involved in political processes. Interviewing often provides a more comprehensive and complicated understanding of phenomena than other forms of research design, and it provides researchers with a rich variety of perspectives.

Conclusion

In this chapter, we discussed two ways of collecting information directly from individuals—through survey research and interviewing. Whether data are collected over the phone, through the mail, on the Internet, or in person, the researcher attempts to elicit information that is consistent, complete, accurate, and instructive. This goal is advanced by being attentive to questionnaire design and taking steps to engage and motivate respondents. The choice of an in-person, telephone, or mail survey can also affect the quality of the data collected. Interviews of elite populations require attention to a special set of issues and generally use a less structured type of interview.

Although you may never conduct an elite interview or a public opinion survey of your own, the information in this chapter should help you evaluate the research of others. Polls, surveys, and interview data have become prevalent in American life. Knowing all the assumptions and judgments that have to be made in carrying out even a small survey may inoculate you from being overly impressed by someone's claim that "the people want" this and that.

[87] Dean and Whyte, "How Do You Know If the Informant Is Telling the Truth?" 127.

[88] Gordon, *Interviewing: Strategy, Techniques, and Tactics,* 18.

[89] Dexter, *Elite and Specialized Interviewing,* 17.

TERMS INTRODUCED

Branching question. A question that sorts respondents into subgroups and directs these subgroups to different parts of the questionnaire.

Closed-ended question. A question with response alternatives provided.

Double-barreled question. A question that is really two questions in one.

Filter question. A question used to screen respondents so that subsequent questions will be asked only of certain respondents for whom the questions are appropriate.

Focused interview. A semistructured or flexible interview schedule used when interviewing elites.

Interviewer bias. The interviewer's influence on the respondent's answers; an example of reactivity.

Interviewing. Interviewing respondents in a nonstandardized, individualized manner.

Leading question. A question that encourages the respondent to choose a particular response.

Open-ended question. A question with no response alternatives provided for the respondent.

Push poll. A poll intended not to collect information but to feed respondents (often) false and damaging information about a candidate or cause.

Questionnaire design. The physical layout and packaging of a questionnaire.

Question-order effect. The effect on responses of question placement within a questionnaire.

Response quality. The extent to which responses provide accurate and complete information.

Response rate. The proportion of respondents selected for participation in a survey who actually participate.

Response set. The pattern of responding to a series of questions in a similar fashion without careful reading of each question.

Sample-population congruence. The degree to which sample subjects represent the population from which they are drawn.

Single-sided question. A question in which the respondent is asked to agree or disagree with a single substantive statement.

Survey instrument. The schedule of questions to be asked of the respondent.

Two-sided question. A question in which two substantive alternatives are provided for the respondent.

SUGGESTED READINGS

Aldridge, Alan, and Kenneth Levine. *Surveying the Social World.* Buckingham, UK: Open University Press, 2001.

Bradburn, Norman, Seymour Sudman, and Brian Wansink. *Asking Questions.* Rev. ed. San Francisco: Jossey-Bass, 2004.

Braverman, Marc T., and Jana Kay Slater. *Advances in Survey Research.* San Francisco: Jossey-Bass, 1998.

Converse, J. M., and Stanley Presser. *Survey Questions: Handcrafting the Standardized Questionnaire.* Beverly Hills, Calif.: Sage, 1986.

Dillman, Don A. *Mail and Electronic Surveys.* New York: Wiley, 1999.

Frey, James H., and Sabine M. Oishi. *How to Conduct Interviews by Telephone and in Person.* Thousand Oaks, Calif.: Sage, 1995.

Nesbary, Dale. *Survey Research and the World Wide Web.* Needham Heights, Mass.: Allyn & Bacon, 1999.

Newman, Isadore, and Keith A. McNeil. *Conducting Survey Research in the Social Sciences.* Lanham, Md.: University Press of America, 1998.

Patten, Mildred L. *Questionnaire Research: A Practical Guide.* 2nd ed. Los Angeles: Pyrczak, 2001.

Rea, Louis M., and Richard A. Parker. *Designing and Conducting Survey Research.* San Francisco: Jossey-Bass, 1997.

Sapsford, Roger. *Survey Research.* Thousand Oaks, Calif.: Sage, 1999.

Tanur, Judith M., ed. *Questions about Questions.* New York: Russell Sage Foundation, 1992.

Weisberg, Herbert F. *The Total Survey Approach.* Chicago: University of Chicago Press, 2005.

CHAPTER 11

Making Sense of Data:

First Steps

CHAPTER OBJECTIVES

11.1 Explain how to compile a data matrix and summarize large batches of data.

11.2 Describe statistics for measuring central tendency and variation or dispersion.

11.3 Relate how to graph data for presentation and exploration.

11.4 Summarize how the early steps in making sense of data lay the groundwork for statistical inference.

MANY STUDENTS WONDER WHY THEY SPEND a semester or more studying statistical methods. After all, aren't topics such as current events, politics in general, law, foreign affairs, voting, and legislatures more interesting? Why bother with something as formal as data collection and analysis? To repeat our sermon in chapter 1, we offer two compelling reasons. First, for better or worse, you need to understand a few basic statistical concepts and methods in order to understand what your—or other people's—numbers mean. Second, good citizenship requires an awareness of statistical concepts. To one degree or another, many issues and policies involve statistical arguments. A story in a widely read St. Louis Web site, for instance, states flatly that "the trend toward greater income inequality has been apparent since the early 1980s—the decade when Gordon Gekko, a fictional character in Oliver Stone's *Wall Street,* first extolled the virtues of greed."[1] Yet conclusions of this sort have

[1] The Editorial Board, "Record Income Inequality Threatens Democracy," *STLToday.com,* September 30, 2010. Available at http://www.stltoday.com/news/opinion/columns/the-platform/article_94a224e8-cce0-11df-a34d-0017a4a78c22.html

been vigorously challenged, especially by conservative economists and journalists. So who's right? Statistical analysis may help.

This chapter takes readers the first steps down the road to understanding applied statistics. Data analysis encompasses three activities: *data exploration, making inferences about hypotheses,* and using the information to *describe and explain* (the term of trade these days is *model*) political phenomena. This chapter covers the first subject; the others are discussed subsequently.

We proceed slowly because the concepts, though not excessively mathematical, do require thought and effort to comprehend. But it will be worth the effort because the knowledge will make you not just a better political science student but also a better citizen.

The Data Matrix

Most of the statistical reports you come across in both the mass media and scholarly publications show only the final results of what has been a long process of gathering, organizing, and analyzing a large body of data. But knowing what goes on behind the scenes is as important as understanding the empirical conclusions. Conceptually, at least, the first step is the arrangement of the observed measurements into a **data matrix**, which is simply an array of rows and columns that stores observed values of variables. Separate rows hold the data for each case or unit of analysis. If you read across one row, you see the specific values that pertain to that case. Each column contains the values on a single variable for all the cases. The column headings list the variable names. To find out where a particular case stands with regard to a particular variable, just look for the row for that case and read across to the appropriate column.

Table 11-1 provides an example, one that pertains to the issue raised in chapters 1 and 2, inequality and power. This matrix contains data for twenty-one developed countries listed in the first column. The Gini index, named after Italian statistician Corrado Gini, measures inequality (in this instance, income inequality). A country with a score of 0 on the index has complete equality in income; everyone has the same income. Higher numbers indicate greater inequality, with a value of 1.0 indicating total inequality; that is, one person has all the income. In the matrix, the Gini measure has been multiplied by 100. Union density is the percentage of employees who are

TABLE 11-1 Inequality and Socioeconomic Measures for Twenty-One Developed Democracies

Country	Gini Index	Union Density (2003)	Social Expenditures (% GDP) (2004)	Percentage of Seats in Legislature Held by "Left-Wing" Parties (2004)	Aged Population (in thousands)	Political Culture
Australia	35.2	23.1	18.777	44.8	2,882	Anglo-American
Austria	29.1	35.7	28.184	56.0	1,283	European
Belgium	33.0	55.6	26.627	60.3	1,790	European
Canada	32.6	28.2	16.569	56.7	4,141	Anglo-American
Denmark	24.7	72.5	27.873	67.9	809	European
Finland	26.9	74.8	25.972	42.4	822	European
France	32.7	8.2	29.452	34.6	9,994	European
Germany	28.3	23.2	27.836	58.5	15,897	European
Greece	34.3	24.5	19.861	49.7	1,989	European
Ireland	34.3	36.3	16.187	56.2	450	Anglo-American
Italy	36.0	34.0	26.151	8.9	10,935	European
Japan	24.9	20.3	18.811	48.7	24,876	—
Luxembourg	30.8	42.3	24.190	51.1	64	European
The Netherlands	30.9	22.4	21.747	54.2	2,270	European
New Zealand	36.2	22.6	18.049	55.4	485	Anglo-American
Norway	25.8	53.0	24.777	28.2	676	European
Spain	34.7	16.2	21.152	47.6	7,186	European
Sweden	25.0	78.0	30.384	66.3	1,548	European
Switzerland	33.7	17.8	27.776	50.2	1,200	European
United Kingdom	36.0	29.2	21.880	59.0	9,570	Anglo-American
United States	40.8	12.6	16.436	46.5	36,301	Anglo-American

Sources: Klaus Armingeon, Romana Careja, Sarah Engler, Panajotis Potolidis, Marlène Gerber, and Philipp Leimgruber, "Comparative Political Data Set III 1990–2008"; Jelle Visser, "Union Membership Statistics in 24 Countries," *Monthly Labor Review* 129, no. 1 (2006), available at http://www.bls.gov/opub/mlr/2006/01/art3abs.htm); Duane Swank, "Electoral, Legislative, and Government Strength of Political Parties by Ideological Group in Capitalist Democracies, 1950–2006: A Database," available at http://www.marquette.edu/polisci/faculty_swank.shtml members of trade unions. In the last column, "Political culture," a rather contrived variable, attempts to capture differences between continental European and British-American political traditions and institutions.

As presented in table 11-1, the data are not very helpful, partly because they overwhelm the eye and partly because it is hard to see even the degree of variability or range of values for the variables, much less what an average value is. (For a much larger data matrix—one with, say, 5,000 rows and 50 variables—the difficulties of interpretation are even worse.) Nor does a matrix reveal many patterns in the data or tell us much about what causes low or high scores. Still, its creation is an essential initial step in data analysis.

Data Description and Exploration

To go from raw data to meaningful conclusions, you begin by summarizing and exploring the information in the matrix. Several kinds of tables, statistics, and graphs can be used for this purpose, but which ones are appropriate to use depends on the level of measurement of the variables. Different statistical procedures assume different levels of measurement. Recall the four broad types of measurement scales:

1. *Nominal:* Variable values are unordered names or labels. (Examples: ethnicity, gender, country of origin)
2. *Ordinal:* Variable values are labels having an implicit but unspecified or measured order. Numbers may be assigned to categories to show ordering or ranking, but strictly speaking, arithmetical operations (e.g., addition) are inappropriate.[2] (Example: scale of ideology)
3. *Interval:* Numbers are assigned to objects so that interval differences are constant across the scale, but there is no true or meaningful zero point. (Examples: temperature,[3] intelligence scores)
4. *Ratio:* In addition to having the properties of interval variables, these scales have a meaningful zero value. (Examples: income, percentage of the population with a high school education)

In this chapter, we clarify which techniques apply to which kinds of variables.

[2] Occasionally, however, it is useful to treat the numbers assigned to the categories of an ordinal or ranking scale as if they were really quantitative.

[3] One metric, the Kelvin scale of temperature, does have an absolute zero, the point at which atoms do not move and heat and energy are absent. The zero points on the other temperature scales are arbitrary.

In the following sections we show how to summarize a large batch of numbers with

- tables (e.g., frequency distributions, cross-tabulations);
- a single number or range of numbers (e.g., mean, maximum, and minimum); and
- graphs (e.g., bar charts).

Frequency Distributions, Proportions, and Percentages

Table 11-2 illustrates a **frequency distribution** of 995 responses to a question regarding the level of influence wealthy people wield in politics: "Do you agree strongly, agree, are uncertain, disagree, or disagree strongly with . . . ? The rich and powerful people in this country have too much influence on politics."

The first column lists the response categories. The second column simply records how many or the *frequency* (often represented by *f*) with which respondents gave each response (e.g., 333 "agree strongly"). More useful indicators are **relative frequencies**, proportions and percentages that help put the raw frequencies into perspective. A *proportion*—the ratio of a part to a whole—is calculated by dividing the number of observations in a category by the total number of observations.

TABLE 11-2 **Frequency Distribution: Beliefs about Power in the United States**

Too Much Influence	Frequency	Proportion	Relative Frequency (%)	Cumulative Frequency (%)
Strongly agree	333	.33	33	33
Agree	533	.54	54	87
Uncertain	38	.04	4	91
Disagree	75	.08	8	99
Disagree strongly	16	.02	2	101
Totals	995	1.01	101	

Question: "Do you agree strongly, agree, are uncertain, disagree, or disagree strongly with . . . ? The rich and powerful people in this country have too much influence on politics."

Source: Marc M. Howard, James L. Gibson, and Dietlind Stolle, *United States Citizenship, Involvement, Democracy (CID) Survey, 2006* (Washington, D.C.: Georgetown University, Center for Democracy and Civil Society [CDACS], 2007). Distributed by Ann Arbor, Mich.: Inter-university Consortium for Political and Social Research ICPSR Study No.: 4607.

HOW IT'S DONE

Proportions and Percentages

Consider a nominal or ordinal variable, Y, with K values or categories and N observations. Let f_k be the frequency or number of observations in the kth category or class. (k goes from 1 to K.)

$\sum_{k=1}^{k}$ means to add frequencies or proportions starting at 1 and stopping at k.

The proportion or relative frequency of cases in the kth category is $\frac{f_k}{N}$.

The **cumulative proportion** in the kth category is

$$\sum_{k=1}^{k} p_1 + p_2 + \dots p_k$$

The percentage of cases in the kth category is

$$\frac{f_k}{N}(100)\cdot$$

A *percentage* is found by multiplying a proportion by 100 or, equivalently, moving the decimal two places to the right. In the third column, which contains proportions, we see that about a third (.33 or 333/995) fall in the first category and about half in the second (.54). In the fourth column these proportions have been converted into percentages. Frequency tables may contain just percentages and not proportions. Finally, the last column, "Cumulative frequency (%)," shows the *cumulative frequencies*. It shows that the overwhelming majority (87.1%) of survey participants either agree strongly or agree with the survey question and thus seem to agree with Hacker and Pierson's assessment that there is an unequal distribution of power in American politics.[4]

MISSING DATA, PERCENTAGES, AND PROPORTIONS. The inclusion or exclusion of invalid and missing values in the total number of observations, which is the base or denominator in the calculation of percentages and proportions, will affect their numerical values and hence our understanding of their substantive meaning. Let's take a look at two examples.

First, look at Table 11-3, based on data from the 2004 National Election Study (NES). One question was, "A working mother can establish just as warm and secure

[4] Jacob S. Hacker and Paul Pierson, "Winner-Take-All Politics: Public Policy, Political Organization, and the Precipitous Rise of Top Incomes in the United States," *Politics & Society* 38, no. 2 (2010): 152–204.

a relationship with her children as a mother who does not work. (Do you agree, neither agree nor disagree, or disagree with this statement)?" Note the sample size, the total of all the cases in the study regardless of whether or not information is available for each and every person, is 1,212. But 152 people out of this total (12.56%) did not offer a substantive response to the question or their responses were for one reason or another not recorded. Therefore, the table includes a subtotal of "valid" or recorded responses. Look at the fourth row, where you will see that there are 1,059 substantive, or valid, responses.

TABLE 11-3 **The Effect of Missing Data on Percentages**

Response	Frequency	Percentage of All Respondents	Percentage of Valid Responses	Cumulative Percentage
Agree	311	25.68	29.37	29.37
Neither agree nor disagree	138	11.40	13.03	42.40
Disagree	610	50.37	57.60	100
Valid responses	1,059	—	100	
Missing	152	12.56		
Total	1,211	100		

Question: "A working mother can establish just as warm and secure a relationship with her children as a mother who does not work. (Do you agree, neither agree nor disagree, or disagree with this statement)?"

Source: National Election Study (Ann Arbor, Mich.: University of Michigan, Center for Political Studies, 2004).

In this case, 29.37 percent *of those (1,059) respondents with substantive or valid responses* agreed that a working woman can establish "just as warm and secure a relationship" with the family as a stay-at-home mom. In the complete dataset, the value was 25.68 percent, a difference of 3.69 percentage points. Here the numbers differ only slightly, but such will not always be the case. The differences in percentages between those with a valid response and the complete dataset may be considerable and might be important.

Imagine someone tells you that a survey proves that 75 percent of Americans favor immigration reform with a pathway to citizenship. You might assume that there must be overwhelming sentiment for this policy. But suppose, as shown in table 11-4, that 1,000 people took part in the poll, but only 200 gave "favor" or

TABLE 11-4 The Effect of Missing Data on Percentages: Large Effect

Response	Frequency	Percentage of All Respondents	Percentage Valid Responses
Favor	150	15	75
Do not favor	50	5	25
Valid responses	200		100
Missing	800	80	
Total	1000	100	

Question: "Do you favor or not favor immigration reform with a pathway to citizenship?"

Source: Hypothetical data.

"do not favor" responses. All the others were recorded as missing—"don't know what a 'pathway' means," "no opinion," or "refused." In this case, concluding that "75 percent" favor immigration reform with a pathway to citizenship might be very misleading because the data also show that only 15 percent of *all* those surveyed favored the policy.

Descriptive Statistics

Frequency distributions, like those displayed in tables 11-2, 11-3, and 11-4, help us make sense of a large body of numbers and consequently are a good first step in describing and exploring the data. They have, however, a couple of shortcomings. First, and perhaps most obvious, it would be nice to have one, two, or at most a few numerical indicators that would in some sense describe the crucial aspects of the information at hand rather than keeping track of many relative frequencies, proportions, or percentages. Another problem with frequency distributions is that they aren't much help in describing quantitative (interval and ratio) variables, for which there is often just one observation for each observed value of the variable. If you refer to table 11-1, for instance, you can see that the twenty-one nations have different Gini scale scores. For these reasons, analysts turn to descriptive statistics.

A **descriptive statistic** is a number that describes certain characteristics or properties of a batch of numbers. In this section, we describe statistics for measuring central tendency and variation or **dispersion**. As with many other statistical procedures we will encounter in later chapters, the appropriate statistic to use depends on the level of measurement.

Measures of Central Tendency

Formally speaking, a measure of **central tendency** locates the middle or center of a distribution, but in a more intuitive sense it describes a typical case. A measure of central tendency applied to table 11-1 can tell you the average or typical Gini coefficient or unionization level of the twenty-one countries shown.

THE MEAN. The most familiar measure of central tendency is the **mean**, called the average in everyday conversation. The mean of a population is denoted by μ. For a sample it is denoted by $\bar{Y}$ (read as "*Y* bar"). A simple device for summarizing a batch of numbers, the mean is calculated by adding the values of a variable and dividing the total by the number of values. For example, if we want the mean of the variable "union density" for the twenty-one developing nations in table 11-1, we just add the values and divide by 21:

$$\bar{Y} = \frac{\begin{gathered}(23.1 + 35.7 + 55.6 + 28.2 + 72.5 + 74.8 + 8.2 + 23.2 + 24.5 + 36.3 + 34.0 + \\ 20.3 + 42.3 + 22.4 + 22.6 + 53.0 + 16.2 + 78.0 + 17.8 + 29.2 + 12.6)\end{gathered}}{21} = 34.79$$

Thus, we can say that, on average, about 35 percent of employees in these countries are unionized.

HOW IT'S DONE

The Mean

The mean is calculated as follows:

$$\bar{Y} \frac{N\sum_{i=1}^{N} Y_i}{N},$$

where *i* refers to the *i* th member of the sample and the symbol $\frac{\sum_{i=1}^{N} Y_i}{N}$, means summing *Y* values starting with

i = 1, *i* = 2, *i* . . . and continuing until all *N* values of *Y* have been added.

The mean is appropriate for interval and ratio (that is, truly quantitative) variables, but it is sometimes applied to ordinal scales in which the categories have been assigned numbers. Everyone uses the mean to get grade point averages (GPAs), which are usually based on the arbitrary practice of assigning a value of 4.0 for an A and so forth. Another substantive example is the mean political ideology as measured in the 2004 NES data mentioned above. The questionnaire asks respondents to place themselves on a 7-point liberalism-conservatism scale, for which the responses are coded 1 for "extremely liberal," 2 for "liberal," 3 for "slightly liberal," 4 for "moderate," 5 for "slightly conservative," 6 for "conservative," and 7 for "extremely conservative." (These integers have no inherent meaning; they are just a way of coding the data.) The mean of the 1,059 cases with valid (that is, nonmissing) values is 4.28. Since the center of the scale, 4, represents a middle-of-the-road position, a mean scale score of 4.28 suggests that the sample is very slightly conservative.

Although the mean is widely known and used, it can mislead the unwary. Here's a simple illustration. Suppose you have been told that Community A has a lower crime rate than Community B. You hypothesize that the gap stems partly from differences in economic well-being. To test this supposition, you take a random sample of ten households per community, obtain the family income of each household, and compute the means for both neighborhoods. The results appear in table 11-5. The mean income of Community A is $37,500; the mean for Community B is $20,500. Since Community A has a higher average (look at the bottom row of the table), you might believe the hypothesis holds water.

TABLE 11-5 **Hypothetical Incomes in Two Communities**

Community A	Community B
$10,000	$10,000
10,000	10,000
12,000	12,000
18,000	18,000
20,000	20,000
22,000	22,000
25,000	25,000
28,000	28,000
30,000	30,000
200,000	30,000
$\bar{Y} = \$37{,}500$	$\bar{Y} = \$20{,}500$

On closer inspection, however, note that the incomes are identical in each community except for the last one. These two families have substantially different earnings. Concentrating on just the mean income of the communities and ignoring any atypical incomes would give you the erroneous impression that people in Community A are financially much better off than people in Community B. In reality, only one family in A is much better off than others in B. This example illustrates how one (or a few) extreme or atypical values can affect or skew the numerical magnitude of the mean (and other statistics). For this reason, other measures of central tendency, known as **resistant measures,** which are not sensitive to one or a few extremes values, are frequently used.

HELPFUL HINTS

The Mean as a Predictor

The mean is used mainly to describe the central tendency of a distribution. Throughout the remaining chapters, however, we put it to a slightly different use. Sometimes it becomes a benchmark for making predictions. Suppose, to take a hypothetical situation, you had a large group of people about whom you knew nothing—not their race, gender, family background . . . nothing. Then you are asked to select a person at random and guess her annual income. Since you have no information, you might feel lost, but you could go to a reference source such as the US Census Bureau. There you might find that in 2013, the mean per capita personal income in the United States was $28,184.[5] In the absence of any other information, a first approximation to the person's yearly income would be $28,184. A bit more formally, this prediction is based on a "model":

Model 1: Predicted income = mean income or, $\hat{Y} = \bar{Y}$.

The little hat over the *Y* means "predicted," and the equation can be read as "the predicted value of *Y* equals the mean of *Y*."

Of course, this method of prediction is going to lead to lots of errors, but mathematics tells us that using the mean to predict the value of a case drawn from a group of numbers is the "best" prediction.

Check out more Helpful Hints at **edge.sagepub.com/johnson8e**

THE MEDIAN. A measure of central tendency that is fully applicable to ordinal as well as interval and ratio data and is resistant to the presence of extreme values is the median. The **median** is a value that divides a distribution in half. That is, *half the observations lie above the median and half below it.*

You can find the middle of an *odd* number of observations by arranging them from lowest to highest and counting the same number of observations from the top and bottom to find the middle. Look at table 11-5. It lists the ordered over-65 population values from the data matrix for twenty-one developed democracies. In this

[5] US Department of Commerce, United States Census Bureau, "Mean Income in the Past 12 Months (in 2013 Inflation-Adjusted Dollars) 2013 American Community Survey 1-Year Estimates." Available at http://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_13_1YR_S1902&prodType=table

example, counting down and up 11 observations will bring you to the middle observation, the 12th.

If you have lots of observations, an easy way to find the middle one is to apply the following formula:

$$mid_{obs} = \frac{(N+1)}{2}$$

For the previous example, this formula yields (21 + 1)/2 = 11, as it should.

If, however, the number of observations is *even*, a modification is required because the middle observation number will contain a 0.5. What to do? Simply use the observations above and below mid_{obs} and average their values. Thus, if we added another country with a "senior" population in thousands as 6,700 we would have 22 countries and the middle observation would be (22 + 1)/2 = 11.5. The middle values would the 11th and 12th, and the median would be the arithmetic mean of the two countries corresponding to those cases: (1,989 + 2,270)/2 = 2,129.50.

The median is a resistant measure in that extreme values (outliers) do not overwhelm its computation. Figure 11-1 shows the calculation of the median for the two hypothetical communities discussed earlier. Recall that the means of the two differed quite a bit: the mean of community A was $37,000 versus $20,500 for community B. But the medians are identical: $21,000. This reflects the fact that the incomes and, hence, the standards of living in the two areas are essentially the same.

TABLE 11-6 **Calculation of the Median**

Over-65 Rank	Population
1	64
2	450
3	485
4	676
5	809
6	822
7	1,200
8	1,283
9	1,548
10	1,790
11 Middle value	1,989
12	2,270
13	2,882
14	4,141
15	7,186
16	9,570
17	9,994
18	10,935
19	15,897
20	24,876
21	36,301

Median = 1,989.

Source: Table 11-1.

THE MODE. A common measure of central tendency, especially for nominal and categorical ordinal data, is the **mode**, or modal category. It is simply the category with the greatest frequency of observations. As an example, start with table 11-2, which shows the distribution of responses to the statement about influence on government. The modal (most frequent) category was "Agree," with 533 responses. It tells us that the modal or typical "belief" about the influence of "the rich and powerful" on government is that they have too much of it. The mode has less utility in describing interval and ratio data, per se, but it is helpful in describing the *shape* of distributions of all kinds of variables. When one category or range of values has many more cases than all the others, we describe the distribution as being *unimodal*, which is to say it has a single peak. But there can be more than one

FIGURE 11-1 **Median Incomes for Communities A and B**

$N = 10$, so $mid_{\text{obs}} = \dfrac{(10 + 1)}{2} = 5.5$. Hence, average the values for the 5th and 6th observations.

	Community A	Community B
1	$10,000	$10,000
2	10,000	10,000
3	12,000	12,000
4	18,000	18,000
5	20,000	20,000
6	22,000	22,000
7	25,000	25,000
8	28,000	28,000
9	30,000	30,000
10	200,000	80,000

$$\longleftarrow M = \frac{(20{,}000 + 22{,}000)}{2} = 21{,}000$$

Source: Table 11-5.

HOW IT'S DONE

The Median

This procedure is practical if the number of cases, *N*, is not large (say, fewer than 30 to 40):

1. Sort the values of the observations from lowest to highest.
2. If the number of cases is an odd number, locate the middle one and record its value. This is the median.
3. If the number of cases is an even number, locate the two middle values. Average these two numbers. The result is the median.

dominant peak or spike in a distribution, in which case we speak of *multimodal* distributions. The term *rectangular* is typically used to describe a distribution that has roughly the same number or proportion of observations in each category. Graphs are often more useful than tables for investigating the "shape" of a distribution, as we will see later.

Of course, summarizing data with a single number has a potential disadvantage because the message contained in one number provides incomplete information. A measure of central tendency does not tell us everything we might like to know about a set of data. For example, suppose a classmate has a GPA of 3.0 (on a 4.0 grade scale). It is impossible to learn from that indicator alone whether she excelled in some courses and struggled in others, or whether she consistently received Bs.

Measures of Variability or Dispersion

We come now to a key concept in statistics: variation, or the differences among the units of a variable. Naturally, we want to know what a typical case in our study looks like. But equally important, we need to take stock of the variability among the cases. The point of many research projects is to understand why this variation arises.

One can get a sense of variability by scanning the values of a variable by reviewing table 11-1. Consider the "Aged population" variable, the data in the sixth column. The figures (in thousands) range from 64 to 36,301, with many values in between. We might conclude that on this variable at least, the nations are quite different or varied or heterogeneous. Now examine "Political culture," a categorical variable. Even a glance reveals that there are only two classes and that most countries in this table have what we call a "European" political culture. Thus, there is not much variation in this variable; rather, the nations are more homogeneous. But this approach gives us only a rough sense of variability, and with many cases it usually is difficult to get a good sense of variability. Measures of variation allow us to express the exact amount of variation in a variable in a single summary number. Ideally, we would use that number to precisely or objectively compare variability among groups of objects or apportion it among known and unknown causes.

PROPERTIES OF MEASURES OF VARIATION. The properties of measures of variation or dispersion described in this book can be summed up in three statements:

1. If there is *no* variability (all the scores have the same value), the measure will equal 0.
2. The measure will always be a positive number (there can't be less than no variation).
3. The greater the variability in the data, the larger the numerical value of the measure.

Many single indices of variation exist, but not all of them have a simple, commonsense interpretation. So it is often helpful to interpret measures of variability with the help of other statistics and graphs.

THE RANGE. For interval- and ratio-level scales, the **range** is a particularly simple measure of variation: it is just the largest (maximum) value of a variable minus the smallest (minimum) value:

$$\text{Range} = \text{maximum} - \text{minimum}.$$

Look carefully at the union densities (union members as percentage of the workforce) in table 11-1. The largest value is 78 percent (Sweden), and the smallest is 8.2 percent (France). The range in union density is therefore 78 – 8.2 or 69.8 percent. In plain language, an enormous disparity exists in labor organization. Now consider spending on "social" programs (e.g., unemployment compensation, pension, and so on) as a percentage of gross domestic product. (See the fourth column of table 11-1.) It varies between a low of 16.2 percent (Ireland) and a high of 30.4 percent (Sweden) for a range of 14.2 percent.

Finally, suppose we had another variable, "level of economic development," coded 1 for developed, 2 for underdeveloped. Surely all the nations listed in table 11-1 would score 1, and there would be no variation. Maximum and minimum values would be 1, and the range (if we treated the codes as actual integers) would be 0. This demonstrates the property that when all units have the same value, a measure of variation (such as the range) will be 0.

INTERQUARTILE RANGE. Another measure of variation, the interquartile range, is easily computed from data that are ordered from smallest to largest. Imagine that, after ordering your data, you divide them into four equal-sized batches. The first bunch would contain 25 percent of the cases, the next would have 25 percent, the next 25 percent, and the last the remaining 25 percent. The division points defining these groups are called quartiles, abbreviated *Q*. Now, find the range as before but use the third and first quartiles as the maximum and minimum. Their difference is the **interquartile range** (IQR):

$$\text{IQR} = Q3 - Q1,$$

where Q3 stands for the third quartile (sometimes it's called the 75th percentile because 75 percent of the cases lie below it) and Q1 is the first (or 25th percentile). Since the interquartile range, a single number, indicates the difference or distance between the third and second quartiles, the middle 50 percent of observations lie between these values.

Another way to think about the computation of the IQR is to obtain the median, which divides the data in half. Then find the medians of *each* half; these medians will be the first and third quartiles.

Let's look at the calculation for the Gini coefficients in table 11-1. Remember, a value close to 0 indicates complete equality, and larger numbers mean greater inequality, with (in this case) 100 signifying maximum inequality. Figure 11-2 shows how the IQR is calculated for the 21 nations in table 11-1. We first find the location of the median: $(N + 1)/2 = (21 + 1)/2 = 11$. Then we find the first quartile by taking the median of the first 11 observations, which is the score for the $(11 + 1)/2 = 6$th case (Germany). Its value is Q1 = 28.3. Similarly, the third quartile is found by calculating the median of the largest 11 values: Q3 = 34.7. The IQR is thus 34.7 – 28.3 = 6.4. We can explain these numbers as saying that three-quarters of the developed countries have Gini scores between about 28 and 35 points. (By contrast, the median for 127 countries across the world is 39.5 and the IQR is 13.5, about twice the IQR for the nations in table 11-1. So we conclude—no surprise—that the world at large is more varied in terms of income inequality than are the industrial democracies.)

FIGURE 11-2 **The Quartiles and Interquartile Range**

Rank	Country	Gini		
1	Denmark	24.7		
2	Japan	24.9		
3	Sweden	25.0		
4	Norway	25.8		
5	Finland	26.9		
6	***Germany***	***28.3***	← Q1 = 28.3	
7	Austria	29.1		
8	Luxembourg	30.8		
9	Netherlands	30.9		
10	Canada	32.6		
11	***France***	***32.7***	← M = 32.7	IQR = 34.7 – 28.3 = 6.4
12	Belgium	33.0		
13	Switzerland	33.7		
14	Greece	34.3		
15	Ireland	34.3		
16	***Spain***	***34.7***	← Q3 = 34.7	
17	Australia	35.2		
18	Italy	36.0		
19	UK	36.0		
20	New Zealand	36.2		
21	USA	40.8		

Quartiles, the range, and the interquartile range (as well as the median) have the property that we have been calling resistance: extreme or outlying values do not distort the picture of the majority of cases. This is a major advantage, especially in small samples. The next set of measures of variability reveal how data diverge from a measure.

DEVIATIONS FROM CENTRAL TENDENCY. We now turn to a different measure of variability. This approach compares departures or deviations from the mean. In table 11-6 we have three sets of data, each containing just three observed values and each having a mean of 50. The last three columns show how much each value deviates from the mean.

The first dataset clearly has no variation, as each value is the same as the mean. The numbers in the second set are almost the same, with deviations of –1, 0, and 1, while those in the third exhibit considerable diversity, with deviations of –30, 0, and 30. The measures of variation to be discussed combine the deviations (the third set of columns) into a single indicator.

THE VARIANCE. The **variance** is the average or *mean* of squared deviations, or the average of all the squared differences between each score and the mean. Denoted σ^2 for a population, the variance can be found by subtracting the mean from each score, squaring the result (a squared deviation), adding up all the squared deviations, and dividing by *N*. The formula for calculating the variance is shown in the box to the right. Note that when calculating the variance for a sample, denoted as s^2 or $\hat{\sigma}^2$, the denominator is adjusted to N–1. Table 11-8 shows the calculation of the sample variance for each of three datasets.

The variance and sample variance follow the rules of a measure of variation: the greater the dispersion of data about the mean, the higher the value of the variance. If all the values are the same, variance equals zero. And it is always nonnegative. The variance is a fundamental concept in mathematical and applied statistics, as we shall see shortly.

TABLE 11-7 Deviations from the Mean

Observed Values			Mean	Deviations from the Mean			Interpretation
50	50	50	50	0	0	0	No deviation implies no variation.
49	50	51	50	–1	0	1	Small deviations imply little variation.
20	50	80	50	–30	0	30	Large deviations imply considerable variation.

Note: Hypothetical data.

HOW IT'S DONE

The Variance for a Sample

The variance is calculated as follows:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{N}(Y_i - \bar{Y})^2}{N-1},$$

where Y_i stands for the ith value, $\bar{Y}$ is the mean of the variable, and N is the sample size.

To simplify calculations, you can use

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{N} Y_i^2 - \frac{\left(\sum_{i=1}^{N} Y_i\right)^2}{N}}{N-1}.$$

THE STANDARD DEVIATION. The most commonly computed and calculated measure of variation is the **standard deviation**, which we denote by σ for a population and by s or $\hat{\sigma}$ for a sample. The standard deviation is simply the square root of the variance. The standard deviation, just like the variance and the mean, is sensitive to extreme values. So notice that in Table 11-8 the last group (49, 50, 80) with its one large value, compared to the other two datasets, has a standard deviation that is more than 17, much larger than for the other datasets.

TABLE 11-8 **Calculation of the Variation and Standard Deviation for Three Hypothetical Sample Populations**

Dataset 1	Squared Deviations	Dataset 2	Squared Deviations	Dataset 3	Squared Deviations
50	$(50-50)^2 = 0$	49	$(49-50)^2 = 1$	49	$(49-59.67)^2 = 113.85$
50	$(50-50)^2 = 0$	50	$(50-50)^2 = 0$	50	$(50-59.67)^2 = 93.51$
50	$(50-50)^2 = 0$	51	$(51-50)^2 = 1$	80	$(80-59.67)^2 = 413.31$
$\bar{Y} = 50$	Sum of squared deviations = 0	$\bar{Y} = 50$	Sum of squared deviations = 2	$\bar{Y} = 59.67$	Sum of squared deviations = 620.67
Variance	$\hat{\sigma}^2 = 0/(3-1) = 0$		$\hat{\sigma}^2 = 2/(3-1) = 1.0$		$\hat{\sigma}^2 = 620.67/(3-1) = 310.34$
Standard deviation	$\hat{\sigma} = \sqrt{\frac{1}{3-1}} = 0$		$\hat{\sigma} = \sqrt{\frac{2}{3-1}} = 1$		$\hat{\sigma} = \sqrt{\frac{620.67}{3-1}} = 17.62$

HOW IT'S DONE

The Standard Deviation for a Sample

The standard deviation is calculated as follows:

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{N}\left(Y_i - \bar{Y}\right)^2}{N-1}},$$

where Y_i stands for the i th value of Y, $\bar{Y}$ is the mean of the variable, and N is the population size.

If you have a calculator that accumulates sums, you can apply this calculating formula:

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{N} Y_i^2 - \frac{\left(\sum_{i=1}^{N} Y_i\right)^2}{N}}{N-1}}.$$

We cannot emphasize enough the importance of carefully exploring data with summary statistics and looking out for cases that may unduly sway the interpretation of the results. The lesson in all of this is that you should not rely on a single number to summarize or describe your data. Rather, use as much information as is reasonable and take your time interpreting each variable. Computers spit out results, but they never interpret them.

MORE ON THE INTERPRETATION OF THE STANDARD DEVIATION. The significance of the standard deviation in statistics is illustrated by considering a common situation. Suppose a large set of data has a distribution approximately like the one shown in figure 11-3. What we see there is a "bell-shaped" distribution called a **normal distribution,** which has the following features:

- The bulk of observations lies in the center, where there is a single peak.
- More specifically, in a normal distribution, half (50 percent) of the observations lie *above* the mean and half lie *below* it.
- The mean, median, and mode have the same numerical values.
- Fewer and fewer observations fall in the tails of the distribution.
- The spread of the distribution is symmetric: one side is the mirror image of the other.

If data have such a distribution, the proportion of cases lying between the mean and a number of standard deviations above and below the mean can be described this way:

- Approximately 68 percent of the data lie between $\bar{Y} - \sigma$ and $\bar{Y} + \sigma$. Read this as "68 percent of the observations are between plus and minus one standard deviation of the mean." For example, if the mean of a variable is 100 and its standard deviation is 10, then about 68 percent of the cases will have scores somewhere between 90 and 110.
- Approximately 95 percent of the cases will fall between $\bar{Y} - 2\sigma$ and $\bar{Y} + 2\sigma$. In the first example, 95 percent or so would be between 80 and 120.
- Almost all of the data will be between $\bar{Y} - 3\sigma$ and $\bar{Y} + 3\sigma$.

This feature of the standard deviation and the normal distribution has an important practical application. For all suitably transformed normal distributions, the areas between the mean and the various distances above and below it, measured in standard deviation units or Z scores, are precisely known and have been tabulated in what is called a "Z table" (see appendix A).

How do we know these percentages? Because mathematical theory proves that normal distributions have this property. Of course, if data are not perfectly normally distributed, the percentages will only be approximations. Yet many naturally

FIGURE 11-3 **Properties of a Normal Distribution**

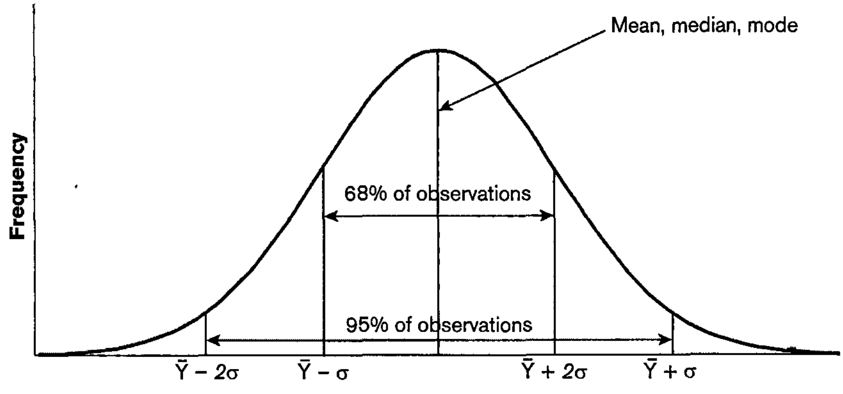


TABLE 11-9 **Summary of Descriptive Statistics**

Statistic	Symbol (if any)	Description (what it shows)	Resistant to Outliers	Most Appropriate for
Measures of central tendency				
Mean	$\bar{Y}$ (for sample) μ (for population)	Arithmetic average: identifies center of distribution	No	Interval, ratio scales
Median	M	Identifies middle value: 50% of observations lie above, 50% below	Yes	Interval, ratio, ordinal scales; ranks
Mode	Mode	Identifies the category (or categories) with highest frequencies	No	Categorical (nominal, ordinal) scales
Measures of variation				
Range	Range	Maximum – minimum	NA	Interval, ratio scales
Interquartile range	IQR	Middle 50% of observations	Yes	Interval, ratio scales
Variance	σ^2 for population s^2 or $\hat{\sigma}^2$ for sample	Average of squared deviations	No	Interval, ratio scales
Standard deviation	σ for population s or $\hat{\sigma}$ for sample	Square root of average of squared deviations	No	Interval, ratio scales

occurring variables do have nearly normal distributions, or they can be transformed into an almost normal distribution.[6] We conclude this section with table 11-9, which summarizes the descriptive statistics discussed in this chapter.

[6] For example, some numbers can be converted to logarithms, which might be normally distributed.

Graphs for Presentation and Exploration

In statistics the maxim "A picture is worth a thousand words" has a special place. We have already discussed the difficulty of using a large data matrix and the need to condense information to a few descriptive numbers. But even those numbers can be uninformative, if not misleading. A major development since the 1980s has been an emphasis on graphical displays to explore and analyze data.[7] These visual tools may lead you to see aspects of the data that are not revealed by tables or a single statistic, and they assist with developing and testing models.

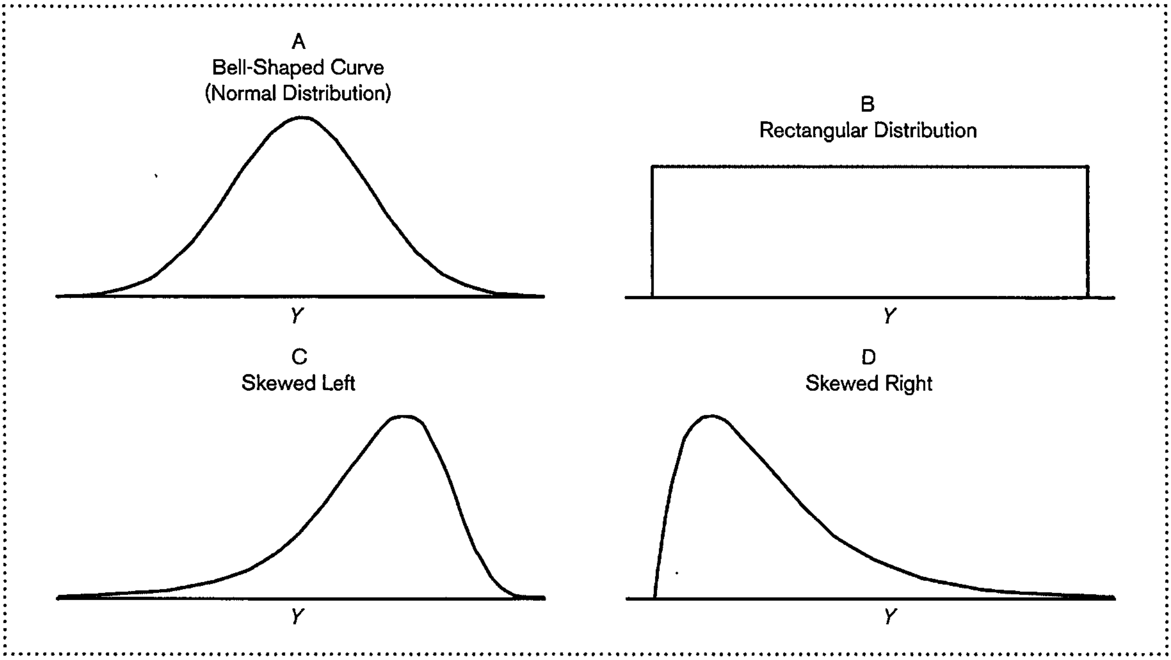
In particular, for a data matrix, a well-constructed graph can answer several questions at one time:

- *Central tendency:* Where does the center of the distribution lie?
- *Dispersion or variation:* How spread out or bunched up are the observations?
- *The shape of the distribution:* Does it have a single peak (one concentration of observations within a relatively narrow range of values) or more than one?
- *Tails:* Approximately what proportion of observations is in the ends of the distribution or in its tails?
- *Symmetry or asymmetry (also called skewness):* Do observations tend to pile up at one end of the measurement scale, with relatively few observations at the other end? Or does each end have roughly the same number of observations?
- *Outliers:* Are there values that, compared with most, seem very large or very small?
- *Comparison:* How does one distribution compare to another in terms of shape, spread, and central tendency?
- *Relationships:* Do values of one variable seem related to those of another?

Figure 11-4 illustrates some ways a variable (*Y*) can be distributed. Panel A displays a symmetric, unimodal (one-peak) distribution, which we previously called bell-shaped or normal. Panel B depicts a rectangular distribution; in this case, each value or range of values has the same number of cases. Panel C shows a distribution that, although unimodal, is **negatively skewed** or skewed to the left. In other words, there are a few observations on the left or low end of the scale, and most observations are in the middle or high end of the scale. Finally, panel D represents

[7] The literature on this topic is vast. For guidelines for presenting accurate and effective visual presentations, Edward Tufte is indispensable. His *The Visual Display of Quantitative Information* (Cheshire, Conn.: Graphics, 1983) is a classic. For an introduction to graphical data exploration, see William Jacoby, *Statistical Graphics for Univariate and Bivariate Data* (Thousand Oaks, Calif.: Sage, 1997).

FIGURE 11-4 **Shapes of Distributions**



the opposite situation: there are comparatively few observations on the right or high end of the scale. The curve is skewed to the right or **positively-skewed**. These are ideal types. No empirical distribution will look exactly like any one of them. Yet if you compare these shapes with the graphs of your own data, you can quickly approximate the kind of distribution you have.

Why is it important to look at the shape of a distribution? For one thing, many statistical procedures assume that variables have an approximately normal distribution. And, if they do not, they can sometimes be mathematically changed or transformed into a bell-shaped configuration.

We can somewhat arbitrarily divide graphs into two general varieties, aimed at two audiences:

1. *Presentation graphs:* Some graphs are intended to be end products. Everyone has seen bar graphs and pie diagrams. The mass media use them to summarize information or even to prove a point, and they commonly appear in books on policy and politics. They are appropriate for summarizing data to be published or otherwise publicly disseminated.

2. *Exploratory graphs:* Graphical analysis works in the background to assist in the exploration of data. As a matter of fact, one of the most vigorous research activities in applied statistics is the search for newer and better ways to visualize quantitative and qualitative information. Beginning with techniques developed in the 1970s by John Tukey and others, research on the visualization of data has become a growth industry.[8] Sometimes these exploratory graphs appear in published research literature, but they are mainly for the benefit of the analyst and are intended to tease out aspects of data that numerical methods do not supply. These diagrams amount to all-in-one devices that simultaneously display various aspects of data, such as central tendency, variation, and shape.

Designing and Drawing Graphs

As we said, a literal picture may be worth a thousand words, but only if it shows what the photographer intends. Computers have greatly simplified the task of constructing graphs with a few keystrokes. Yet this computational power will create problems if overused. Remember, a graph is supposed to provide the viewer with a visual description of the data. Computers may churn out wonderful-looking charts. But many of them add so many extra features, such as three-dimensional bars, exploded slices, cute little icons, or colorful fills, that the data easily get lost in the ink. Edward Tufte called these doodads "chartjunk."[9] It is usually best to keep lines and areas as simple as possible so that readers can easily see the point being conveyed. More important, the essence of successful graphing is to ensure that the viewer can accurately perceive the data and any relations among variables.[10] The image should not distort the data. The graphical elements (e.g., size of plotting symbols, colors, axes, etc.) should reflect the actual relationships and trends in the data. Graph designers introduced the term *lie factor* to quantify this notion:[11]

$$\text{Lie factor} = \frac{\text{Size of effect displayed in graph}}{\text{Size of effect in data}}.$$

Many statisticians and social scientists regard a graph as a story about the data. Like any story, it should be well told and not drone on and on.

[8] John Tukey, *Exploratory Data Analysis* (New York: Addison-Wesley, 1977).

[9] Tufte, *The Visual Display of Quantitative Information,* 107–21.

[10] Kevin J. Keen, *Graphics for Statistics and Data Analysis with R* (Boca Raton, Fla.: CRC Press, 2010), 11. Chapter 1 of this book succinctly describes the "principles of statistical graphics."

[11] Cited in ibid., 18.

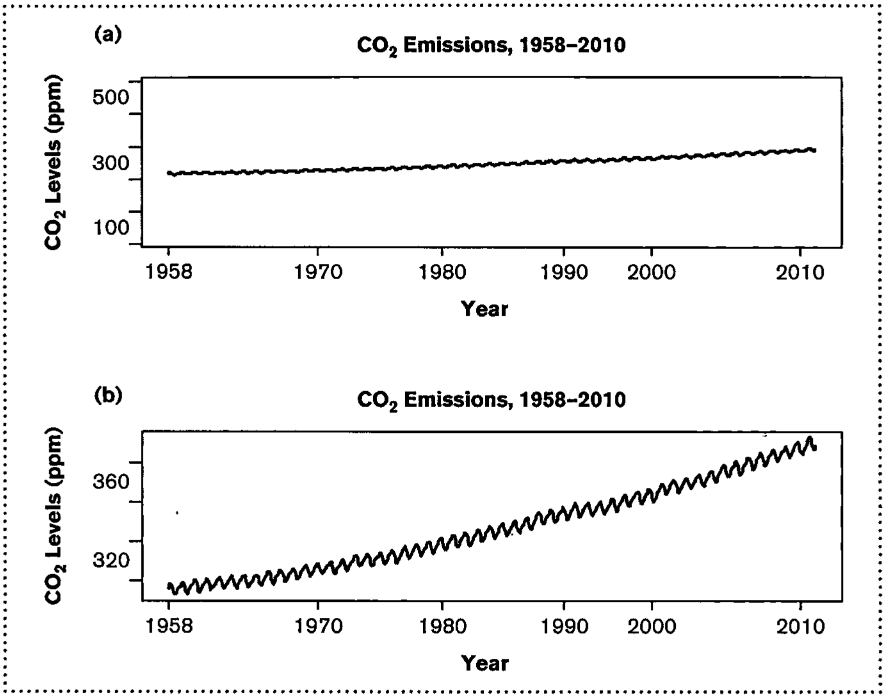
Here are some tips. They are very general. Their purpose is merely to alert you to the fact that unless helped by humans, computers are not especially good at telling stories.

- Small amounts of data (few cases) usually don't need to be summarized by graph. A table (e.g., frequency distribution) is often a better way to present small amounts of numerical information.
- A table that spans several pages, however, may overwhelm a reader unless it's meant to be part of a databook. But graphs can provide a succinct summary, especially if one wants to emphasize trends or patterns and looking at individual data points is not essential to the story.
- Pick an appropriate type of graph. We see shortly that categorical (nominal and ordinal) variables require or are best suited to one variety, quantitative variables to another. A mixture of variable types (one nominal and one continuous, for example) may require special attention.
- Think carefully about the axes and how they relate to the scale of the data. There are two considerations: the range of data values and how they are measured on the graph and the physical dimensions of the plot. Suppose, as we do shortly, you want plot carbon dioxide (CO_2) emissions against year. If the time axis is 5.5 inches wide, the increase over the years may look relatively flat; but if the axis is only 2 inches, the rates may seem to soar. Consider the data's maximum and minimums before setting the axes' limit(s). Both axes should be proportional to the data's limits. The viewer should have little problem understanding the range. Scale the axes to show trends in a reasonable way (e.g., figure 11-5.) Is there a meaningful zero point? Including it may or may not be appropriate.
- Clearly label axes and graphical elements (e.g., bars, lines, symbols). Note the measurement units in labels, titles, and legends (e.g., "Population [in thousands]," "Spending [as percent of GDP]").
- Use sufficient tick or "hash" marks so the reader can quickly estimate quantities.
- Don't "clip" (not include) extreme values that are part of the data unless absolutely necessary and clearly noted. It is sometimes necessary to cut an axis, but make sure everyone knows what you are doing and why.
- If possible, identify interesting or extreme values, but if the graph contains a lot of points (more than, say, 30 or 40), labeling all of them may leave an unreadable mess on the page. Use text (sparingly) to point out interesting features of the data. Suppose a plot of GDP per capita versus total expenditures on health (as a fraction of GDP) reveals a clear pattern (the greater the wealth, the greater the spending), but one country lies out of the mainstream. It would be worth identifying this country on the graph.

- Independent variables usually (but not always) appear on the *x*- or horizontal axis. A common exception is when comparing categories of nominal or ordinal factors in terms of a quantitative variable. Then analysts frequently list the classes along the *y*- or vertical axis and extend bars or boxes into the graph (from left to right) to show magnitudes along the horizontal scale.
- Again, the size of the graphical elements has to be proportional to the data they represent. Only use the size of a graphical element, such as a bar or circle, to show differences in data; keep these elements the same size otherwise. Suppose you are using bars to show the percentages of a sample that are liberal and conservative. If the liberal bar is twice as wide as the one for conservatives (1 inch versus 1/2 inch), the relative proportions may be distorted in the viewer's eye, even if the actual numbers appear in the graph. Politicians along with statisticians know that the visual often dominates the verbal.
- If you are using bars to represent data, it helps to arrange them in a logical order. If the variable is ordinal, then list the categories from lowest to highest. Similarly, group common categories together. Typically in charts that show the relative proportions of individuals who self-identify with one or the other US political party, strong to weak Democrats are placed next to each other, as are Republicans, with independents placed in the middle. Imagine what the picture would look like if the categories were randomly scattered.
- Avoid 3-D effects like the plague unless they help tell your story. Of course, a three-dimensional graph may be necessary for a multiple-variable display, but normally not to dramatize or beautify. Avoid unneeded decorations.
- Always include a title.
- Indicate the source and date of the data if possible.
- Points are usually better than icons.
- Colors are useful to show different categories, etc., but as in this book, the graph has to stand on its own in black and white. As Web page designers are well aware, your viewer may not have the same viewing device you used to create the image. And don't use more colors than data values.
- A rule of thumb: when it comes to graphing, the less ink on the page, the better.

As an example of some of these issues, consider the debate about global warming. Most climatologists argue that dramatic increases in carbon dioxide (CO_2)—a "greenhouse gas"—emissions have risen so steeply as a result of human activities that the Earth's average temperature is rising to the point where human well-being is threatened. Let's lay aside the pros and cons of the argument and instead look

FIGURE 11-5 CO_2 Emissions, 1958–2010: Dramatic Climate Change?



Source: "Global Greenhouse Gas Reference Network," Earth System Research Laboratory. Accessed June 22, 2015. Available at http://www.esrl.noaa.gov/gmd/ccgg/trends/#mlo_data

at how one can use a graph to inform and misinform citizens. Figure 11-5 shows the trend in CO_2 emissions, in parts per million (ppm), over the past half century or so. Consider two ordinary citizens. One who looks at only (a) might say, "Why shucks, the levels haven't changed hardly t'all. What's the worry?" The second, however, sees only (b) and thinks, "Oh my God! The levels are skyrocketing." Why the different reactions? Exactly the same data have been plotted, but the *y*-axes are different. In (a), the limits are 0 and 500. Yet the maximum and minimum of the series are 313 and 393, respectively, a spread of 80 ppm. This difference gets swallowed in the *y*-axis, making the trend look almost level. In (b) the *y*-limits are now the minimum and maximum values.[12] The graph seems to surge from the bottom and crash into the top of the frame. Neither graph is flat-out lying. It's just that the

[12] Needless to say, the graphs are only suggestive. The next step is to attempt to "model" the data. This task lies beyond the scope of the book. But we should note that trends are not easy to analyze because they can be the result of real changes caused by some set of factors, or of random fluctuations, or both. You might be amazed at how a process can drift randomly and yet appear to be a true increase or decrease.

length of the *x*-axis (compared to the *y*-scale) and the different limits of the *y*-scale create differing visual impressions, especially among casual viewers. Although this sort of squished presentation commonly appears in journals, it might be preferable to use a larger plotting region. In other respects, though, the graphs are relatively concise, informative, and clean.

Bar Charts

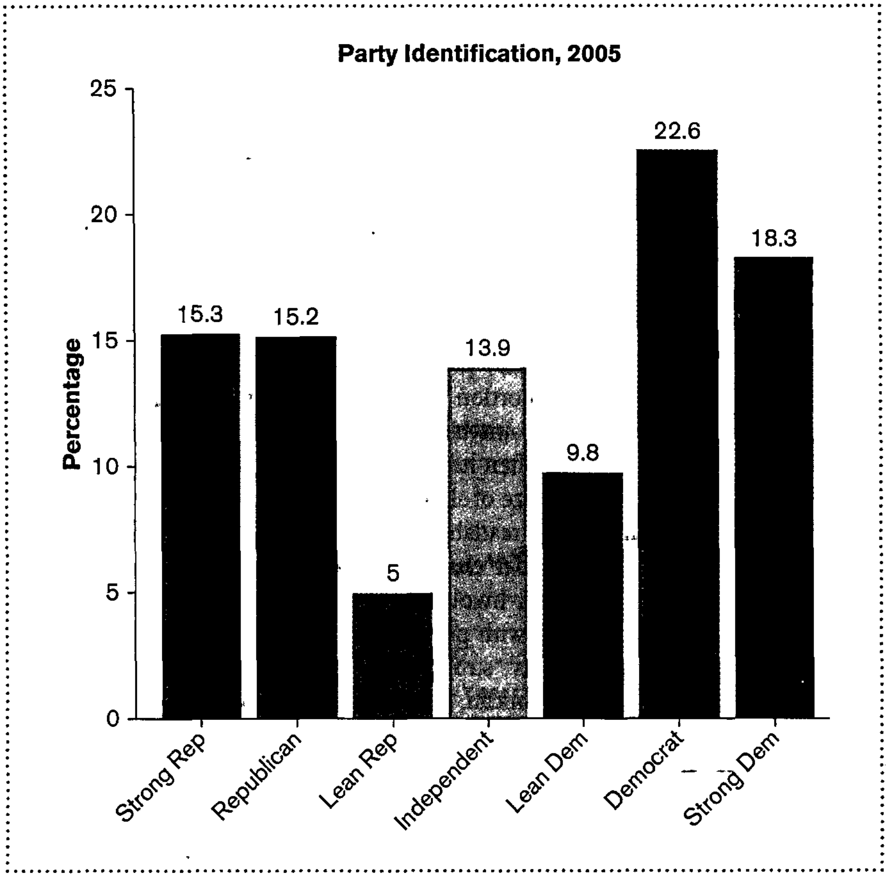
Numerical information can be presented in many ways, the most common of which are bar charts and pie diagrams. A **bar chart** is a series of bars in which the height of each bar represents the proportion or percentage of observations that are in the category. A **pie chart** is a circular representation of a set of observed values in which the entire circle (or pie) stands for all the observed values and each slice of the pie is the proportion or percentage of the observed values in each category. Pie charts or diagrams have fallen into disfavor among many statisticians partly because viewers often have difficulty making accurate assessments and comparisons of the relative size of slices and partly because bar charts provide the same information in a more readable form. Hence, we do not discuss them further. Figure 11-6 presents a bar chart of the party identification data found by the United States Citizenship, Involvement, Democracy (CID) Survey, 2006, mentioned earlier in connection with partisanship (see table 11-2). This time we plot party identification coded as "strong Democrat," "Democrat," "lean Democrat," independent," "lean Republican," "Republican," and "strong Republican." The bar chart tells the same story: Democrat is the modal category, and there are slightly more self-identified Democrats than Republicans, relatively few "leaners" (in this sample), a roughly rectangular distribution (most categories, except the leaning Republicans, differ by not much more than 8 percent). In a presentation, you would include one chart *or* the other.

These types of graphs are most useful when the number of categories or values of a variable is relatively small and the number of cases is large. They most commonly display percentages or proportions. Be careful about constructing a bar chart when you have, say, a dozen or more categories. By the same token, these graphs will not reveal much if you have fewer than ten or fifteen cases. And you have to make sure the graphical elements (bars) are proportional to the data. Unless there is a substantive reason to do otherwise, keep miscellaneous and missing value categories out of the picture.

Exploratory Graphs

The graphs in this section display empirical frequency distribution for different types of variables. They can be presented in formal reports but are often

FIGURE 11-6 Bar Chart of Party Identification



Source: United States Citizenship, Involvement, Democracy (CID) Survey, 2006.

used behind the scenes to explore the properties of a batch of numbers in one picture.

DOT CHART. A dot chart displays *all* of the observed values in a batch of numbers as dots along a horizontal or vertical line that represents the variable. Since it shows the entire dataset, the number of observations should be relatively small—for example, fewer than fifty. The great advantage of this plot is that it presents the main features of a distribution. To construct a simple dot chart, draw a horizontal line that stands for the variable. Below the line, print a reasonable number of values of the variable. Finally, using the data scores to position the dots, draw one dot above the line for each observation.

HELPFUL HINTS

Inspect Graphs First

Participants in debates frequently use graphs to bolster their arguments. In most instances, it's a good idea to scan a graph before reading the author's claims about what it says. The application of data analysis to real-world problems is at least as much a matter of good judgment as it is an exact science. A researcher, who will have a lot invested in demonstrating a point, may see great significance in a result, whereas your independent opinion might be "big deal." You can maintain this independence by first drawing your own conclusions from the visual evidence and then checking them against the writer's assessments. If you don't study the information for yourself, you become a captive, not a critic, of the research.

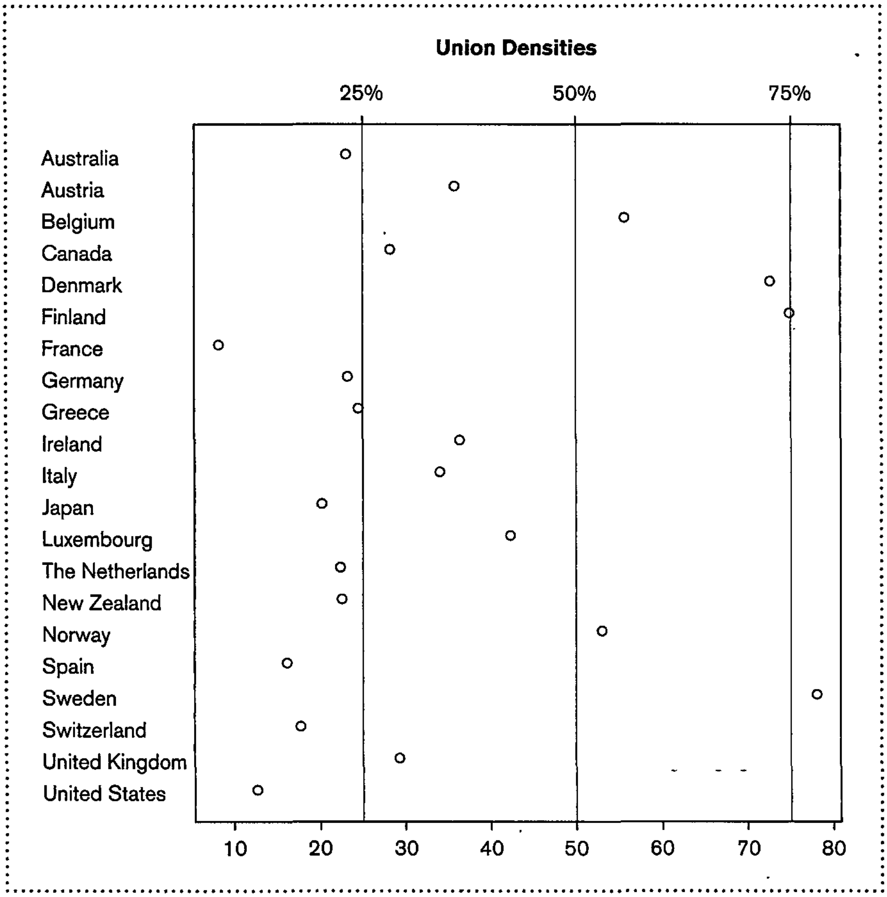
Check out more Helpful Hints at **edge.sagepub.com/johnson8e**

Figure 11-7 shows a simple dot chart of union density rates in the developed countries introduced in table 11-1. It shows at a glance which countries have low and high densities. The United States and France (perhaps surprisingly) have the lowest levels, while only two (Sweden and Finland) are 75 percent or above. The mean and median appear to be between 30 and 50 percent. (Actually the median is 28.2 percent, and the mean is 34.8 percent.) We see clearly that the two "outlying" Scandinavian countries (Sweden and Finland) pull the median down from the mean. This result, in turn, tells us that the distribution is skewed slightly toward the lower end of the scale.

In view of widespread and widely publicized labor protests in France, we might wonder if we have made a recording error for France (8 percent). Consequently, it is necessary to check the recorded value against the original data. As it happens, the percentage is approximately correct, but according to a European Community publication, "in membership terms the French trade union movement is one of the weakest in Europe with only 8% of employees in unions. . . . But despite low membership and apparent division French trade unions have strong support in elections for employee representatives and are able to mobilise French workers to great effect."[13] The case shows that a graph such as a dot chart can reveal situations

[13] Worker-Participation.eu, "Trade Union." Available at http://www.worker-participation.eu/National-Industrial-Relations/Countries/France/Trade-Union/

FIGURE 11-7 Dot Chart



Source: Table 11-1.

that might warrant further investigation, not only to check the data's accuracy but also to explain the apparent anomaly. And it would also be worth pondering why, among all these industrial nations, the United States has such a comparatively anemic labor movement.

The dot chart is actually quite versatile, and depending on available software, it can display combinations of categorical factors against a quantitative dependent variable. (In this simple example, the independent variable is "country," and the category labels are just the country names.) It is common practice, for example, to sort the values of the dependent variable from lowest to highest and display the ordered data. Or one can use symbols and colors to highlight important features.

HISTOGRAMS. A **histogram** is a type of bar graph in which the height and area of the bars are proportional to the frequencies in each category of a nominal or ordinal variable or of a continuous variable in intervals. If the variable is continuous, such as age or income, construct a histogram by dividing the variable into intervals, or bins, counting the number of cases in each one, and then drawing the bars to reveal the proportion of total cases falling in each bin. If the variable is nominal or ordinal with a relatively few discrete categories, just draw bar sizes so as to reflect the proportion of cases in each class.

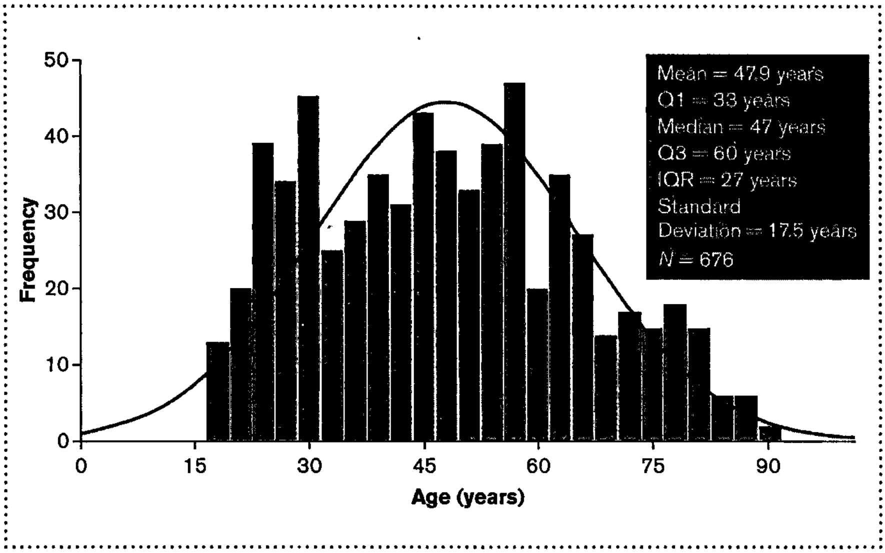
A histogram, like other descriptive graphical devices, reduces a data matrix in such a way that you can easily see important features of the variable. For example, you can quickly see modes if there are any, the shape of the distribution (that is, whether or not it is skewed), and even extreme values. It helps to annotate your exploratory graphs with summary statistics so that you have everything required for analysis in one place.

Figure 11-8 shows a histogram of women's ages from the 2004 National Election Study. (Several midpoints of the intervals are shown on the *x*-axis.) It is important to examine this variable because (1) we want to make sure the sample distribution approximates the population, and (2) if discrepancies exist, we can identify and adjust for them. Although a couple of spikes can be observed in this distribution, the overall appearance is very roughly normal or bell-shaped. The graph succinctly sums up the more than 600 female respondents in the sample. Notice among other things that the middle of the distribution, as measured by both the mean and the median, is just about 47. The lower end of the scale is truncated because females less than 18 years of age were ineligible for participation in the study. The "location" statistics (Q1, median, and Q3) are instructive: 50 percent of the women are somewhere between 33 and 60 years old, and fully 25 percent are older than 60 years.

If you have to construct a histogram by hand (unlikely in this computer-dominated era), draw a horizontal line and divide it into equal-sized intervals or bins for the variable. The *y*-axis shows the frequency or proportion of observations in each interval. Simply count the number of units in each group and draw a bar proportionate to its share of the total. Suppose you have a total of 200 cases, of which 20 fall in the first interval. The bar above this interval should show that 20 observations, or 20/200 = 10 percent of the observations, are in it.

Most statistical programs, even elementary ones, easily draw histograms. This capability comes in handy because an investigator may want to try creating histograms on a given dataset using many different numbers of intervals. Histograms are helpful, as we have indicated, because they summarize both the spread of the values and their average magnitude. They are, however, quite sensitive to the delineations or definitions of the cut points or bins. By *sensitive,* we mean that the shape of the

FIGURE 11-8 Histogram of Women's Ages (Normal Curve for Comparison)



Source: 2004 National Election Study.

distribution can be affected by the number and the width of the intervals. Some programs do not give the user much control over the intervals, so be cautious when using them.

BOXPLOT. Perhaps the most useful graphical device for summarizing and exploring interval- and ratio-level data is the boxplot. It does not display individual points but does explicitly and simultaneously let you see several descriptive statistics: Q1, the median, Q3, and what are called the "lower" and "upper" whiskers, which for now can be thought of as fences beyond which any data points far from the "ordinary" lie. (We discuss these a bit more later.) What is more, it can be annotated in various ways to reveal even more information. Boxplots are sometimes called box-and-whisker plots because they appear to have a whisker at each end of a box.

Constructing a boxplot (even with paper and pencil) is relatively simple:

1. Find the maximum and minimum, the first and third quartiles, the interquartile range (IQR), and the median.
2. Draw a horizontal line to indicate the scale of the variable. Mark off intervals of the variable. Be sure to fully label the scale.

3. Above the line, say about half an inch or so, draw a small vertical line to indicate the median. It should correspond to the appropriate value on the scale.

4. Next, draw short vertical lines of the same size above the scale to indicate Q1 and Q3.

5. Sketch a rectangle with the two quartiles (Q1 and Q3) at the ends. The median will be in the box somewhere. The height of the rectangle does not matter.

6. Next, calculate 1.5 times the interquartile range, IQR.

7. Calculate the "lower whisker." The lower whisker is the *maximum* of either (1) the minimum of the variable, *or* (2) 1.5 times the IQR. In symbols, the lower whisker equals the maximum of (minimum [variable], 1.5 x IQR). (It looks complicated, but in reality you get used to doing this pretty quickly, and most statistical software does the work automatically.) Call this quantity "LW" for short.

8. Draw a line a distance LW from the left end (Q1) of the box.

9. Do the same for the "upper whisker." This time, however, you take the *minimum* of either (1) the maximum of the variable, *or* (2) 1.5 times the IQR. More succinctly, the upper whisker is the lesser of the maximum value of the variable, or 1.5 x IQR. Call the result "UW."

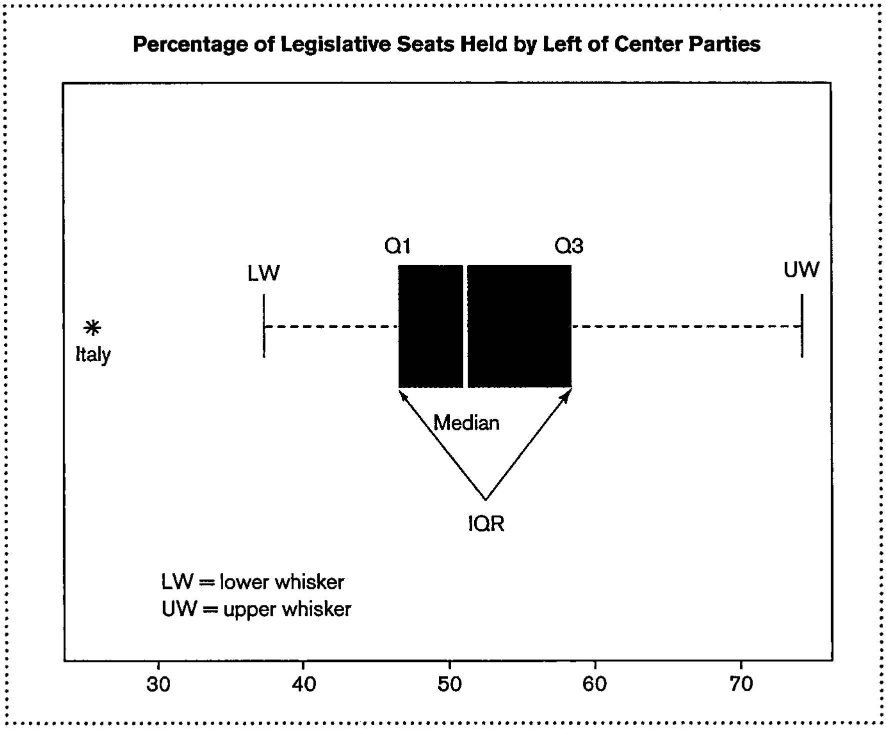
10. Draw a line from the third quartile (Q3) to the point UW.

11. Place points or symbols to indicate the actual location of extreme values. These should be labeled with the observation name or number.

12. Give the graph a title and properly label the *x*-axis.

This may seem complicated, so let's look at an example, this time using a different variable: "percent of legislative seats held by leftist (e.g., Democrats in America) parties." (See table 11-1; the data are from 2004.) In later chapters, we will invoke this variable as a possible causal factor affecting social-welfare expenditures and equality, as Hacker and Pierson's study suggests. The boxplot in figure 11-9 gives us a bird's-eye view of the distribution. The bottom scale is measured in percentages. The solid line in the middle of the box represents the median, which we can see is slightly more than 50 percent. (The actual number is 52.1 percent.) The lines at the end of the box, marked Q1 and Q3, are the first and third quartiles. (The precise boundaries of the box are 46.5 and 58.5.) They show that in about 75 percent of these nations, left-wing parties held between roughly 47 and 59 percent of the seats in the legislatures in 2004. The lower and upper whiskers show the location of extreme values. The boxplot has been annotated to show its main features. In

FIGURE 11-9 **Boxplot of Left-of-Center Strength in Legislatures**



Source: Table 11-1.

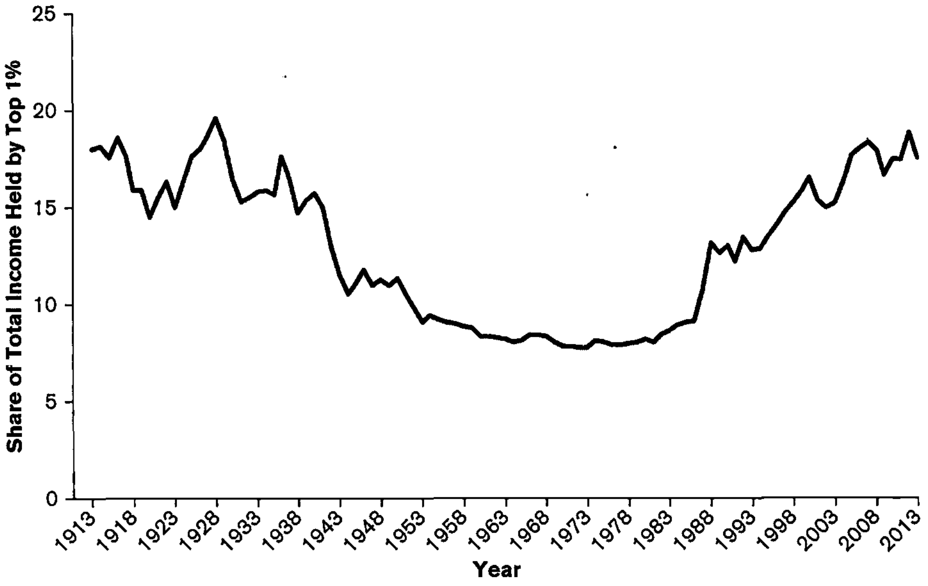
most circumstances, you wouldn't bother with these explanatory notes. That is, the notations on the graph ("median," "Q1," "Q3," and so forth) are not usually included because viewers supposedly understand this kind of graph. We include them merely for instructional purposes. We now have a graphical summary of the "left party strength" variable.

TIME SERIES PLOT. Political scientists frequently work with variables that have been measured or collected at different points in time. In a time series plot, the *x*-axis marks off time periods or dates, whereas the *y*-axis represents the variable. These sorts of plots frequently appear in newspapers and magazines and could well be described as a type of presentation graph. However, they are also helpful in exploring trends in data in preparation for explaining or modeling them.

The time series plot in figure 11-10 shows the share of total income held by the wealthiest 1 percent of Americans each year from 1913 to 2013. As you can see, the rich had about 18 percent in 1913; that portion dropped to 7.74 percent in 1973 and then began rising sharply after the late 1970s. By 2012 the percentage had increased to close to 20. It is data like these that convince many analysts that economic inequality is on the increase in the United States. One can use the picture to explain the fluctuations. What was there about the period from 1950 to 1975, say, that led to less concentration at the top? And what might explain the reversal after 1978? If you know recent history, or are familiar with relevant literature, you might note that at this time foreign competition accelerated, thereby costing thousands of jobs in America and lower wages. Or, since the late 1970s there has been a growth in both business power and conservatism. These factors might account for the seeming growth in inequality.

Table 11-10 summarizes the kinds of graphs we have discussed and offers a few tips on their proper use.

FIGURE 11-10 **Top 1% Income Share in the United States, 1913–2013**



Source: The World Top Incomes Database, http://topincomes.parisschoolofeconomics.eu/#Database

TABLE 11-10 **Typical Presentation and Exploratory Graphs**

Type of Graph	What Is Displayed	Most Appropriate Level of Measurement	Number of Cases	Comments
Bar chart	Relative frequencies (percentages, proportions)	Categorical (nominal, ordinal)	3-10 categories	Common presentation graphic
Dot chart	Frequencies, distribution shape, outliers	Quantitative (interval, ratio)	*Less than* 50 cases	Displays actual data values
Histogram	Distribution shape	Quantitative	$N > 50$ cases	Essential exploratory graph for interval or ratio variables with a large number of cases
Boxplot	Distribution shape, summary statistics, outliers	Quantitative	$N > 50$ cases	Can display several distributions; actual data points, an essential exploratory tool
Time series plot	Trends	Quantitative (percentages, rates)	$10 < N < 100$	Common in presentation and exploratory graphics

Note: Entries are guidelines, not hard-and-fast rules.

What's Next?

So far we have described and explored our datasets with various numbers, tables, and graphs. If we have carried out our research competently, we have the beginnings of a quantitative analysis that will eventually lead (we expect) to some answers to our substantive questions (what explains cross-national variation in inequality or welfare spending or why some people become more active in politics than others). If, as in the case of the comparative data of table 11-1, we have a complete set of cases (countries), we can proceed to a more thorough analysis of the hypotheses. On the other hand, think about the survey or poll data we touched on (e.g., the National Election Study). This material comes from a *sample* and thus raises a question: Are the results a reflection or indicator of reality, or do they merely represent chance and would differ considerably if we drew another sample? In other words, can we infer that what we observe in a sample accurately reflects the true situation for the population as a whole? How accurately? These are the questions addressed by statistical inference to which we now turn.

TERMS INTRODUCED

Bar chart. A graphical display of the data in a frequency or percentage distribution.

Central tendency. The most frequent, middle, or central value in a frequency distribution.

Cumulative proportion. The total proportion of observations at or below a value in a frequency distribution.

Data matrix. An array of rows and columns that stores the values of a set of variables for all the cases in a dataset.

Descriptive statistic. A number that, because of its definition and formula, describes certain characteristics or properties of a batch of numbers.

Dispersion. The distribution of data values around the most frequent, middle, or central value.

Frequency distribution. The number of observations per value or category of a variable.

Histogram. A type of bar graph in which the height and area of the bars are proportional to the frequencies in each category of a nominal variable or intervals of a continuous variable.

Interquartile range. Difference between third and first quartiles.

Mean. The sum of the values of a variable divided by the number of values.

Median. The category or value above and below which one-half of the observations lie.

Mode. The category with the greatest frequency of observations.

Negatively skewed. A distribution of values in which fewer observations lie to the left of the middle value and those observations are fairly distant from the mean.

Normal distribution. A distribution defined by a mathematical formula and the graph of which has a symmetrical, bell shape; in which the mean, the mode, and the median coincide; and in which a fixed proportion of observations lies between the mean and any distance from the mean measured in terms of the standard deviation.

Pie chart. A circular representation of data in which the entire circle (or pie) stands for all the observed values and each slice the proportion or percentage of observations in each category.

Positively skewed. A distribution of values in which fewer observations lie to the right of the middle value and those observations are fairly distant from the mean.

Range. The distance between the highest and lowest values or the range of categories into which observations fall.

Relative frequency. Percentage or proportion of total number of observations in a frequency distribution that have a particular value.

Resistant measure. A measure of central tendency that is not sensitive to one or a few extreme values in a distribution.

Standard deviation. A measure of dispersion of data points about the mean for interval- and ratio-level data.

Variance. A measure of dispersion of data points about the mean for interval- and ratio-level data.

SUGGESTED READINGS

Abelson, Robert P. *Statistics as Principled Argument.* Hillsdale, N.Y.: Lawrence Erlbaum, 1995.

Agresti, Alan. *An Introduction to Categorical Data Analysis.* New York: Wiley, 1996.

Agresti, Alan, and Barbara Finlay. *Statistical Methods for Social Sciences.* Upper Saddle River, N.J.: Prentice Hall, 1997.

Cleveland, William S. *Visualizing Data. Summit,* N.J.: Hobart Press, 1993.

Jacoby, William. *Statistical Graphics for Univariate and Bivariate Data.* Thousand Oaks, Calif.: Sage, 1997.

Lewis-Beck, Michael. *Data Analysis: An Introduction.* Thousand Oaks, Calif.: Sage, 1995.

Tufte, Edward R. *Beautiful Evidence.* Cheshire, Conn.: Graphics, 2006.

———. *The Visual Display of Quantitative Information.* 2nd ed. Cheshire, Conn.: Graphics, 2001.

Velleman, Paul, and David Hoaglin. *Applications, Basics, and Computing of Exploratory Data Analysis.* Pacific Grove, Calif.: Duxbury Press, 1983.

CHAPTER 12 Statistical Inference

CHAPTER OBJECTIVES

12.1 Discuss two core activities of statistical inference.

12.2 Explain confidence intervals and confidence levels.

DATA OBTAINED FROM THE AMERICAN NATIONAL Election Study, 2012, show that on a feeling "thermometer" scale, which runs from 0 for very cool to 100 for very warm, "strong Democrats" give "illegal immigrants" an average rating of 50; "strong Republicans," by contrast, rate this group at about 25.[1] Given these data, one might conclude that Democrats are much more favorably disposed to undocumented individuals than are Republicans. Yet these estimates, 50 and 25, are based on samples.[2] How do we know for sure that strong Democrats and Republicans differ on this issue in the total population? Is this difference the result of sampling error? (See chapter 7.)

Answering these and similar questions that arise whenever political scientists use samples rather than populations to measure political phenomena brings

[1] These data were obtained from the "2012 ANES Time Series," available from computer-assisted survey methods program (csm) at the University of California, Berkeley, "SDA: Survey Documentation and Analysis." Accessed January 14, 2015. Available at http://sda.berkeley.edu/sdaweb/analysis/?dataset=nes2012

[2] In this instance the samples exceed 800 cases per group.

us to the topic of statistical inference. Statistical inference helps investigators decide which results or effects occurred by happenstance and which are manifestations of reality.

Two Kinds of Inferential Activities

Statistical inference means many things to many people, but for us it involves two core activities:

1. *Hypothesis testing:* Many empirical claims can be translated into specific statements—hypotheses—about a population, which can be confirmed or disconfirmed with the aid of probability theory. So we might hypothesize that there is no difference in opinion between strong Democrats and strong Republicans.
2. *Point and interval estimation:* The goal here is to estimate unknown population parameters from samples and to surround those estimates with confidence intervals. Confidence intervals suggest the estimate's reliability or precision. Using our example, we might be able to say, "We are 95 percent sure that the difference in the mean thermometer rating of illegal immigrants between strong Democrats and strong Republicans is 25 points plus or minus 3 percent."

We discuss each activity in turn.

Hypothesis Testing

Statements called **statistical hypotheses** are key to hypothesis testing. There are two types: null hypotheses and research or alternative hypotheses.

Null hypotheses have two important characteristics: They are succinct and precise assertions about population parameters, such as a mean equals a certain value, a pair of proportions does not differ, or a numerical indicator of a relationship between two variables is zero. In many research reports, the null hypothesis (H_0) is that something (for example, a mean or a proportion) equals zero. Hence, the word *null*—because zero represents no effect, such as no difference. But keep in mind that a null hypothesis can be an assertion that a population parameter equals any *single* number such as 0.5 or 100.

> There is often more than one way to state a null hypothesis. So, for example, if we use D to represent the difference in mean approval ratings, the null hypothesis would be
>
> H_0: D = 0.
>
> Alternatively, the null hypothesis could state that the mean (μ) approval ratings were the same for strong Democrats and strong Republicans:
>
> $H_0: \mu_{SR} = \mu_{SD.}$

2. They are stated in such a manner that data plus statistical theory allow us to reject them with a known degree of confidence that we are not making a mistake.

In addition to stating a null hypothesis, researchers state another hypothesis called the **research or alternative hypothesis**, represented by H_A. Researchers usually hope that they will be able to reject the null hypothesis in favor of their research hypothesis. Again, there are several ways an alternative hypothesis might be stated. If the research hypothesis was specific and indicated the direction of the expected difference, e.g., strong Republicans have a lower approval rating of illegal immigrants compared to strong Democrats, then the alternative hypothesis could be stated:

$$H_A: \mu_{SR} < \mu_{SD}.$$

If a researcher is unable to specify the direction of a relationship, then the alternative hypothesis could be stated:

$$H_A: \mu_{SR} \neq \mu_{SD} \text{ or } D \neq 0.$$

Obviously, the null and alternative hypotheses cannot both be true. Deciding which to accept and which to reject involves the possibility of making a mistake or error.

TYPES OF ERRORS IN STATISTICAL INFERENCE. To paraphrase dictionary definitions, inference refers to reasoning from available information or facts to reach a conclusion. The end product, however, is not guaranteed to be true. It might in fact turn out to be incorrect. This possibility certainly arises in the case of statistical inference, in which values of unknown population characteristics such as means or proportions are estimated and hypotheses about those characteristics are tested.

In hypothesis testing—that is, making a decision about a null hypothesis—two kinds of mistakes are possible, as illustrated in table 12-1. The first type of mistake is to reject a true null hypothesis. Statisticians call this mistake a **type I error**. The probability of making a type I error is normally designated by the lowercase Greek letter alpha, α. Another possible mistake is failing to reject a null hypothesis that is false. This type of error is called a **type II error**. The probability of committing a type II error is normally designated with the lowercase Greek letter beta, β.[3]

TABLE 12-1 **Types of Inferential Errors**

	In the "Real" World, the Null Hypothesis Is . . .	
Decision is to	True	False
Accept H_0	Correct decision	Type II error
Reject H_0	Type I error	Correct decision

The convention for testing hypotheses is to focus on the probability of making a type I error. Tests of statistical significance calculate this probability. When researchers claim that their results are "statistically significant," they are claiming that the null hypothesis has been rejected with a specified probability of making a type I error.

LEVELS OF SIGNIFICANCE. The term *level of statistical significance* is used to refer to the probability of making a type I error. The three most common levels of **statistical significance** in political science are .05, .01, and .001, but it is really up to the researcher (and you) how great a chance to take of making a type I error, and you do not need to be bound by these conventions. In many situations, the actual value of p, typically used to represent the probability of making a type I error, is reported leaving the decision up to the reader. Yet you are likely to encounter a statement such as "The result is significant at the .05 level." The statement means that a researcher has set up a null hypothesis, drawn a sample, calculated a sample statistic or estimator, and found that its particular value would occur by chance at most only 5 percent of the time, if the null hypothesis is true.

Steps for Hypothesis Testing

The process for testing statistical hypotheses involves several steps. Let's review these before we work with some examples.

1. State a null hypothesis, e.g., H_0: $\mu = 80$.
2. State an alternative hypothesis, e.g., H_A: $\mu \neq 80$, or $\mu > 80$, or $\mu < 80$ (choose only one of these alternatives).

[3] The probability of making a type II error depends on how far the true value of the population parameter is from the hypothesized one. In addition for a fixed α, the probability of a type II error (β) decreases as the sample size increases. The probability of detecting and thus rejecting a false null hypothesis is called the power of the test and equals $1 - \beta$. Power is an extremely important issue in statistics. Many commonly used inferential tests may have relatively low power. Excellent introductions to the topic are Jacob Cohen, *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. (Hillsdale, N.J.: Erlbaum, 1988); and Jacob Cohen, "A Power Primer," *Psychological Bulletin* 112, no. 1 (1992): 155–59, available at http://www.math.unm.edu/~schrader/biostat/bi02/Spr06/cohen.pdf

3. Make a decision rule. This involves deciding the maximum probability of making a type I error you are willing to accept (level of statistical significance). This level of statistical significance is represented by α and called the "alpha" level.
4. Determine whether you are conducting a "one-tailed" or a "two-tailed" test of statistical significance. If you are predicting that your sample statistic is different from the population parameter, but you are not predicting whether it is smaller or larger, you will be conducting a two-tailed test. If you are predicting that the sample statistic is larger (or smaller) than the population parameter, you will be conducting a one-tailed test.
5. Choose an appropriate test statistic and sampling distribution. The appropriate test statistic and sampling distribution will depend on the type of population parameter you are estimating and the size of your sample. Remember, a sampling distribution is a mathematical function that indicates the probability of different values of the sample statistic or estimator occurring.[4] Think of a sampling distribution as a picture of the variability of sample results.
6. Determine the critical value of the test statistic. The critical value is the value of the test statistic that must be obtained in order to reject the null hypothesis at the specified alpha level.
7. Calculate the sample statistic or estimator of the population parameter.
8. Calculate the observed value of the test statistic. The general formula for calculating this value is:

$$\text{Observed test statistic} = \frac{(\text{Sample estimate} - \text{Hypothesized populatation parameter})}{\text{Estimated standard error of sample statistic}}.$$

9. Compare the observed value of the test statistic to the critical value of the test statistic.

The decision to reject or not reject the null hypothesis depends on the comparison between the observed test statistic and the critical value.

- Two-tailed test: Reject H_0 if the absolute value of the observed test statistic is greater than or equal to the critical value.
- One-tailed test: Check to make sure the sample estimate is consistent with H_A. If so, reject H_0 if the absolute value of the observed test statistic is greater than or equal to the critical value.

[4] Building statistical inference on the idea of repeated samples initially makes students uneasy since it is indeed a difficult concept. Even more interesting is that it bothers many researchers in the field. For a readable introduction to this debate, see Bruce Western, "Bayesian Analysis for Sociologists," *Sociological Methods and Research* 28, no. 1 (1999): 7–11.

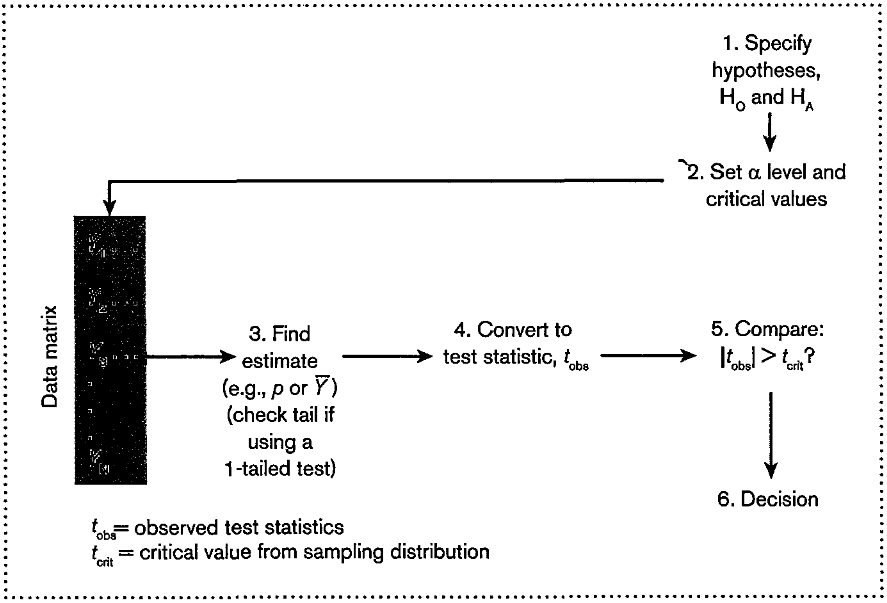
Figure 12-1 summarizes the steps in hypothesis testing. Now, let's see how the process of hypothesis testing works with some examples.

Significance Tests of a Mean

Suppose someone tells you the "average American has left the middle of the road and now tends to be somewhat conservative." You, however, are not so sure. You think that the average American is *not* conservative. This situation could be tested using responses to a 7-point ideology scale with 1 representing "extremely liberal," 4 "moderate" or "neither liberal nor conservative," and 7 "extremely conservative." A 5 on this scale represents a slightly conservative position. So the above claim can be interpreted as saying that the mean ideology score of voting-age individuals is 5. The null hypothesis is, thus, H_0: $\mu = 5$, where μ is the population mean ideology score. Given the way the problem has been set up, the alternative hypothesis is H_A: $\mu < 5$, which indicates that you are hypothesizing the true value to be something closer to middle of the road or perhaps even liberal. For this example, let's set $\alpha = .05$ for the level of significance.

The sample statistic for this test is the sample mean of liberalism-conservatism scores. The next step is to specify an appropriate sampling distribution that determines which test statistic you will be using and gather the data for the test. When

FIGURE 12-1 **Steps in Hypothesis Testing**


1. Specify hypotheses, H_O and H_A
2. Set α level and critical values
Data matrix
3. Find estimate (e.g., p or $\bar{Y}$) (check tail if using a 1-tailed test)
4. Convert to test statistic, t_{obs}
5. Compare: $|t_{obs}| > t_{crit}$?
6. Decision
t_{obs}= observed test statistics
t_{crit} = critical value from sampling distribution

testing a hypothesis about a mean, you will use the sample size to determine the appropriate sampling distribution.

LARGE-SAMPLE TEST OF A MEAN. An important theorem in statistics states that, given random samples of size 30 or more cases, the distribution of the sample means of a variable (*Y*) is approximately a normal distribution with a mean equal to μ (the mean of the population from which the sample was drawn) and a standard deviation of $\hat{\sigma}_{\bar{Y}} = \hat{\sigma} / \sqrt{N}.$, also known as the *standard error* of the mean. It measures how much variation (or imprecision) there is in the sample estimator of the mean. For us, the theorem boils down to the fact that we can test large-sample means with a standard normal distribution.

The graph of the standard normal distribution is a unimodal, symmetrical (bell-shaped) curve. This particular distribution has a mean of zero and a standard deviation of 1. (Figure 11-3 displays a graph of the standard normal distribution.) The areas between the mean and any point along the horizontal axis have been tabulated. (Appendix A contains such a table.) So if you travel up from the mean a certain distance, you can easily find how much of the area under the curve lies beyond that point. Thinking of these areas as probabilities, you can thus establish critical regions and critical values. The values along the horizontal scale, called ***z* scores**, are multiples of the standard deviation. For example, $z = 1.96$ means 1.96 standard deviations above the mean. Remember that a standard error of a mean is a standard deviation of a sample mean, so we can also think in terms of a *z* score as a multiple of a standard error as well.

HELPFUL HINTS

Summary of Notation

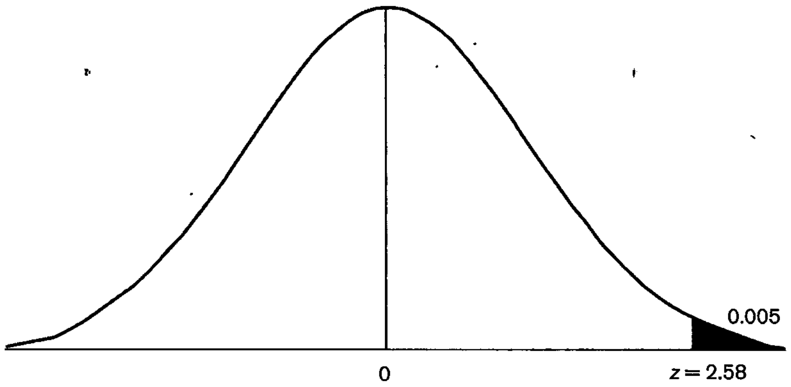
In the text the symbols have these meanings:

- Y: the variable of interest
- $\bar{Y}$: the sample mean
- $\hat{\sigma}_Y$: the sample standard deviation
- $\hat{\sigma}_{\bar{Y}}$: the estimated standard error of $\bar{Y}$ (Note the bar over the Y.)

Check out more Helpful Hints at **edge.sagepub.com/johnson8e**

Figure 12-2 shows you how to use the tabulated distribution to find critical values. Suppose we want to conduct a one-tailed test of a hypothesis at the .005 level. The numbers appearing in the body of the table show the area or proportion of the distribution lying above the *z* scores defined by the row and column headings. For example, scan down the column under "z" until you come to the row marked "2.5." This corresponds to a *z* value of 2.5. Now move across the row until you come to the entry ".0049." Then move up the column to the top row under "Second Decimal Place of *z*." There you should see ".08." The combination of the row label (2.5) and column label (.08) gives 2.58 (just add 2.5 and .08). The area at and above this *z* score is .0049 or about .005. That's the size of the region we want, so the critical value for the test is 2.58. In probability language, 2.58 creates a critical region for which (assuming H_0 is true) the probability of a sample result's landing in it is .005.

FIGURE 12-2 Using the Standard Normal Distribution



Normal Curve Tail Probabilities
Excerpted from Appendix A
Second Decimal Place of *z*

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
...	...	...	...	...	...	...	...	...	...	...
2.5	.0026	.0600	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
2.9	.0019	.0018	.0017	.0017	.0016	.0016	.0015	.0015	.0014	.0014
...	...	...	...	...	...	...	...	...	...	...

HOW IT'S DONE

Calculating the Observed *z*

The observed *z* is calculated as follows:

$$z_{\text{obs}} = \frac{(\bar{Y} - \mu)}{\hat{\sigma}/\sqrt{N}},$$

where $\bar{Y}$ is the sample mean, μ is the hypothesized population mean, $\hat{\sigma}$ is the sample standard deviation, and *N* is the sample size.

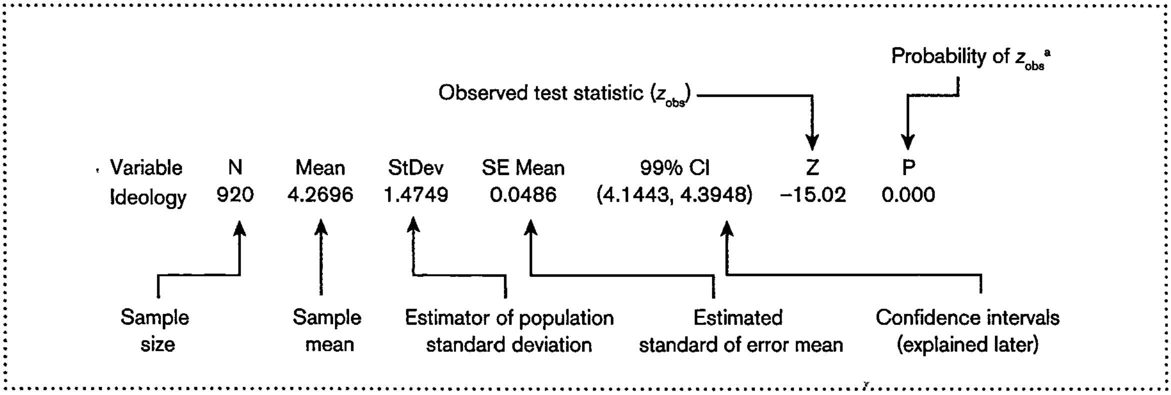
ONE-TAILED TEST OF A SAMPLE MEAN. Let's put the *z* distribution to work testing our hypothesis. Using 2004 NES data, we find that $\bar{Y} = 4.27$, the sample size is 920, and the standard deviation of the sample is 1.47. Because we are conducting a one-tailed test, we must first check to see if $\bar{Y}$ differs from the null hypothesized value of 5 in the direction we have indicated in our alternative hypothesis. Since 4.27 is less than 5, so far so good. But is 4.27 enough lower than 5 so that we can reject the null hypothesis with α set to .05? To answer this question, we must next find the critical value for *z*. Using the *z* table in appendix A, we find that the critical value for our alpha level is 1.65 (the area above this *z* score is .0495, or about .05). Now we must calculate the observed *z* so we can compare it to the critical value of *z*. In order to reject the null hypothesis, the absolute value of the observed *z* must be equal to or greater than the critical value. The observed *z* is calculated as follows:

$$z_{\text{obs}} = \frac{(\bar{Y} - \mu)}{\frac{\hat{\sigma}}{\sqrt{N}}} = \frac{(4.27 - 5)}{\frac{1.47}{\sqrt{920}}} = \frac{-.73}{.048} = -15.21.$$

The absolute value of this result greatly exceeds the chosen critical value (1.65), so the null hypothesis would be rejected. As a matter of fact, the observed *z* exceeds *any* value in the tabulated standard normal distribution. Consequently, we would conclude that the probability of making a type I error is vanishingly small. Figure 12-3 shows how most software represents the probability (as 0.000). This does not mean that there is *no* possibility the null hypothesis is true; it only suggests a very small likelihood that cannot be presented conveniently.

What are we to make of this highly statistically significant result? It could be presented with great fanfare. We might declare that, based on our statistical evidence, Americans can in no way be construed as being slightly conservative. But the sample mean is 4.27, a value ever so slightly in the conservative direction. And how much

FIGURE 12-3 **Results from a Large-Sample Test of a Mean (Example *z*-Test Results from a Software Package)**



```
Variable    N     Mean    StDev   SE Mean        99% CI            Z        P
Ideology  920   4.2696   1.4749    0.0486   (4.1443, 4.3948)   -15.02   0.000
```

Probability of z_{obs}[a]
Observed test statistic (z_{obs})
Sample size
Sample mean
Estimator of population standard deviation
Estimated standard of error mean
Confidence intervals (explained later)

[a]The probability of observing a z statistic this large or larger under the null hypothesis that $\mu = 5$.

does a somewhat arbitrary ideology scale reveal about attitudes, and how much substantive or practical importance can we place on scale score differences of 0.5 or even 1.0? The soundest conclusion seems to be that the public is in the middle of the political road, or just a bit to the right.

TWO-TAILED TEST OF A SAMPLE MEAN. Using the same 2004 NES data, let's conduct the test at the .01 level and use a two-sided test, so now the alternative hypothesis is H_A: $\mu \neq 5$. We need two critical regions, but their total area must equal .01 to give the desired level of significance. This means that the size of each tail must be .01 / 2 = .005. We know from our earlier exploration of the z table that the critical value is 2.58. The table gives values only for the upper half of the distribution, but the normal distribution is symmetric, so –2.58 (note the minus sign) defines an area at the lower end equal to about .005. Consequently, if we get an observed test statistic that is greater than or equal to either –2.58 *or* +2.58, we will reject the null hypothesis at the .01 level. Our observed z of –15.21 hasn't changed and it still greatly exceeds –2.58, so we are again quite safe in rejecting the null hypothesis. There is a very small probability that we are incorrectly rejecting a true null hypothesis.

To help secure the procedure in your mind, let's find the critical values for (1) a two-tailed test at the .05 level, and (2) a one-tailed test at the .002 level.

1. Since we want a two-tailed test, we have to divide .05 in half and look for the z value in the table that marks off the .05/2 = .025 proportion of the distribution. Look in appendix A for ".0250," the size of the critical

region. When you have found it, look at the row and column labels (you may want to use a straight edge). The row should be "1.9" and the column ".06" so that the critical value is 1.9 +.06 = 1.96. This value is compared to the observed z to arrive at a decision.

2. Check to make sure the sample mean differs from the hypothesized population mean in accordance with H_A. We need be concerned with only one end of the distribution. Therefore, search the bottom of the table for ".002." Again, when it has been located, the conjunction of row and column labels should be "2.88." You would compare this number to the observed z to make the decision.

Table 12-2 summarizes the test criteria, sample values, and decision for testing the hypothesis that the average population liberalism-conservatism score is 5 using a two-tailed test.

SMALL-SAMPLE TEST OF A MEAN. If the sample is small—less than or equal to about 30—statistical theory asserts that the appropriate sampling distribution for a test about a mean is the *t distribution*.[5] A graph of this distribution is a symmetrical (bell-shaped) curve with a single peak. It resembles the normal distribution, but is a bit "fatter" in that it has more area in its tails. The shape of the t distribution depends on the sample size (N) and is thus a "family" of distributions. But as N gets larger, the t distribution approaches the shape of the normal distribution; at N = 30 or 40, they are essentially indistinguishable.

TABLE 12-2 Large-Sample 2-Tailed Test of a Mean

Null hypothesis	H_0: $\mu = 5$
Alternative hypothesis	H_A: $\mu \neq 5$
Sampling distribution	Standard normal
Level of significance	$\alpha = .01$
Size of each critical region	.005
Critical value	$z_{crit} = 2.58$
Sample size	920
Sample mean ($\bar{Y}$)	4.27
Estimated population standard deviation ($\hat{\sigma}$)	1.47
Estimated standard error ($\hat{\sigma}_{\bar{Y}}$)	.048
Observed test statistic	$z_{obs} = -15.21$

To use the t distribution for a small-sample hypothesis test of a mean, follow these steps:

1. Choose the level of significance and directionality of the test, a one- or two-tailed test at the α level.
2. Find the degrees of freedom (*df*) by calculating $N - 1$. The degrees of

[5] The t table in appendix B gives t values for sample sizes up to 120, but once the sample size reaches 30, the difference between the t and z distributions is small. It is generally acceptable practice to use the z distribution for samples of size 30 or above, but it is also not uncommon for researchers to continue to use the t distribution for samples larger than 30.

freedom is for now an abstract concep that we do not explain. But in this situation it is always the sample size minus 1.

3. Given these choices, find the critical value(s). Looking at a tabulation of a t distribution (see appendix B for an example), go down the rows of the table until you locate the degrees of freedom. Move across the row until you find the column that corresponds to the area of the size of the designated level of significance. The number at the intersection of the degrees of freedom row and area under the curve column is the critical value, t_{crit}.

4. Calculate the observed t value. The observed t is found by using the same formula we used for calculating an observed z:

$$t_{\text{obs}} = \frac{(\bar{Y} - \mu)}{\hat{\sigma} / \sqrt{N}}.$$

5. If $|t_{\text{obs}}| \geq t_{\text{crit}}$, reject the null hypothesis.

Let's test the statistical hypothesis that the mean liberalism-conservatism score for the population is 5. Suppose we use a small sample of 25 observations from the 2004 National Election Study (NES) and we set $\alpha = .05$ as before. Figure 12-4 shows how the critical value is found from the t distribution shown in appendix B. (For the sake of brevity, many table entries have been deleted.)

The level of significance, .05, for a one-tailed test is found by looking in the second column. The degrees of freedom for this problem is calculated as $25 - 1 = 24$, so we use the 24th row. The intersection of this row and the third column leads to the critical value, 1.711. Therefore, if the observed test statistic equals or is greater than 1.711, we reject the null hypothesis in favor of the alternative. Otherwise, we do not reject.

The mean liberalism-conservatism score for our sample of 25 turned out to be 4.44, with a standard deviation of 1.23. This observed mean is slightly below the hypothesized average. Thus, our sample estimate is consistent with the alternative hypothesis. To make an obvious point, if the sample mean had been 5 or higher, we would proceed no further as we are conducting a one-tailed test and such an outcome would not allow us to reject the null hypothesis. In this case, we can proceed and calculate the observed t:

$$t_{\text{obs}} = \frac{(\bar{Y} - \mu)}{\frac{\hat{\sigma}}{\sqrt{N}}} = \frac{(4.44 - 5)}{\frac{1.23}{\sqrt{25}}} = \frac{-.56}{.246} = -2.28.$$

FIGURE 12-4 **Finding the *t* Value**

Level of significance = total size of critical region for one-sided test.
(Probability of type I error (α level)

	Alpha Level for One-Tailed Test						
	.05	.025	.01	.005	.0025	.001	.005
	Alpha Level for Two-Tailed Test						
Degrees of Freedom (*df*)	.10	.05	.02	.01	.005	.002	.001
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
21	1.721	2.080	2.518	2.831	3.135	3.527	3.819
22	1.717	2.074	2.508	2.819	3.119	3.505	3.792
23	1.714	2.069	2.500	2.807	3.104	3.485	3.767
24	1.711	1.064	2.492	2.797	3.091	3.467	3.745
25	1.708	2.060	2.485	2.787	3.078	3.450	3.725
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .

df = 25 − 1 = 24

Critical value for α = .05 with 24 *df*.

Source: Excerpt from "Critical Values from *t* Distribution," appendix table B, p. 606.

Notice that the test statistic can be negative, but we are only interested in its absolute value (that is, disregarding the negative sign). And since $|t_{obs}|$ = 2.28 is greater than 1.711 (the critical value), we reject the null hypothesis. The implication is that Americans are not as conservative as hypothesized, and at this point the best estimate of the true mean is 4.44, a very slightly conservative value.

To cement your understanding of this section, use the sample information provided and the decision rule to test the hypothesis that H_0: $\mu = 4$. We have supplied all the necessary information. Then, for further practice, pick a different decision criterion, say, $\alpha = .01$ for the level of significance.

Computer programs now perform most statistical analyses. Although the advantages in saved time and effort are obvious, it is essential to understand what the computer output or a table in a research article is telling us. That's why we invest so much time in going over the ideas behind hypothesis testing. But whether as a student or in another capacity, you are likely to be a consumer of software-generated reports. Figure 12-5 illustrates the results of a small-sample *t*-test cranked out by a popular software package. Instead of indicating that the result is significant at, say, the .05 level, this program, like most others of its kind, gives the probability of getting a *t* statistic *at least as large as the one actually observed if the null hypothesis* *is true*. In the present case, in which the sample mean is 4.44, the evidence is that a population value of 5 is not very likely. More precisely, the probability of a sample mean this far or farther from the hypothesized value is only about .016 or 16 chances in 1000.

Finally, given the ubiquity of computers, you might as well take advantage of their services and follow this rule: whenever the *p*-value is available, report it and not an arbitrary level of significance. Why follow this advice? Compare these two assertions:

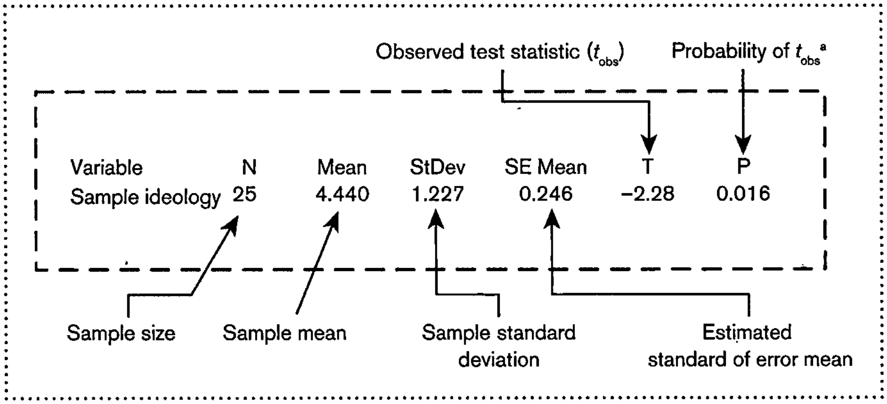
1. The result is significant at the .05 level.
2. The *p*-value is .016.

The first statement tells us only that the probability of the result (or one more extreme) is less than .05. But is it .04 or .02? The second statement indicates that under the null hypothesis the probability of the result (or one more extreme) is .016; this statement is more specific.

Testing Hypotheses about Proportions

Throughout this book, and indeed throughout political science, we everywhere come across proportions and their first cousins, percentages. One can form a

FIGURE 12-5 Results from a Small-Sample Test of a Mean (Example *t*-Test Results from a Software Package)



```
Variable          N    Mean    StDev   SE Mean    T       P
Sample ideology  25    4.440   1.227   0.246    -2.28   0.016
```

Observed test statistic (t_{obs})
Probability of t_{obs}[a]
Sample size
Sample mean
Sample standard deviation
Estimated standard of error mean

Source: See table 11-1.

[a] The probability of observing a *t* statistic this large or larger under the null hypothesis that $\mu = 5$.

statistical hypothesis about them as well. Suppose, for example, we want to estimate the proportion of citizens who donate money to political organizations and causes. After asking our question of a randomly drawn sample, we could record the responses as 0 for "No" and 1 for "Yes." After the surveys have been tallied, we could find the average, which would just be the total of the scores divided by the sample size:

$$\frac{(0+0+0+0+\ldots+1+1+1+1\ldots)}{N}=\frac{\mathit{Total\ Number\ of\ 1s}}{N}=p,$$

which as we see it is just the proportion of respondents coded 1 or the proportion saying yes. We treat $\hat{p}$ as a sample estimator of the population proportion, P. The test that P equals a particular value follows the same procedure used for the mean:

1. State null and alternative hypotheses.
2. Determine the sampling distribution of p.
3. Decide on a decision rule (α level, test direction, critical value).
4. Obtain the data and calculate the estimator and its sample standard error.
5. Compare this result to the chosen critical value.
6. Decide whether or not to reject the null hypothesis.

As with the mean, sample distributions of proportions depend on N: for small samples (<30), the t distribution is appropriate, while for larger samples, the z or standard normal distribution takes over. Assuming the truth of the null hypothesis, the sampling distribution will have a mean P and a standard error (deviation) $\hat{\sigma}_P$:

$$\hat{\sigma}_p=\sqrt{\frac{p(1-p)}{N-1}}=\sqrt{\frac{pq}{N-1}},$$

where $\hat{p}$ is the sample proportion, $1-p=q$ is the proportion *not* in the category of interest, and N is the sample size. We use this to find the observed test statistic in the usual way:

$$\text{Observed test statistic}=\frac{\text{Sample proportion}-\text{Hypothesized proportion}}{\text{Standard error of proportion}}=\frac{p-P}{\hat{\sigma}_p}=t_{\text{obs}}\ or\ z_{\text{obs}}.$$

We use data from the United States Citizenship, Involvement, Democracy study to gauge the level of political contributions. The questionnaire asked its participants, "During the last 12 months, have you done any of the following? Donated money to a political organization or group." As suggested earlier, those saying "no" were coded 0; those saying "yes" received scores of 1. The estimated proportion turns out to be 202/995 = 0.203. (There were 995 valid responses out of which 202 had made political donations.) Given this particular sample, $\hat{p} = .2$ is our best estimate of the true level of public giving. Now, suppose we believed on the basis of small pilot surveys and other information that the real proportion is .3. Do we have reason to believe that our sample result differs (meaningfully) from the hypothesized value, $P = .3$?

Since the true value could be larger or smaller than .3, the alternative hypothesis is that P is either less or greater than .3, which leads to a two-tailed test. This "test of a proportion" has the usual layout, with the main difference being the statistic of interest is now a proportion. We calculate the standard error as

$$\hat{\sigma}_p = \sqrt{\frac{(.20)(1-.2)}{995-1}} = \sqrt{\frac{.16}{994}} = .0128.$$

So the test statistic is

$$z_{\text{obs}} = \frac{(p - P)}{\hat{\sigma}_p} = \frac{(.2030 - .3)}{.0128} \approx -7.578.$$

(We have used more precision for the calculations than are reported in the text.) Table 12-3 summarizes the steps.

A small-sample test of a proportion follows the same procedure but uses the t distribution in place of the standard normal distribution.

TABLE 12-3 **Large-Sample Test of a Proportion**

Null hypothesis	H_0: $P = .3$
Alternative hypothesis	H_A: $P \neq .3$
Sample size	995
Sample statistic	Sample proportion
Sampling distribution	Standard normal (z)
Level of significance	$\alpha = .01$
Size of each critical region	.005
Critical value	$z_{\text{crit}} = 2.57$
Sample proportion of "yes"	0.203
Estimated population standard deviation ($\hat{\sigma}$)	.40
Estimated standard error ($\hat{\sigma}_p$)	.0128
Observed test statistic	$z_{\text{obs}} = -7.578$

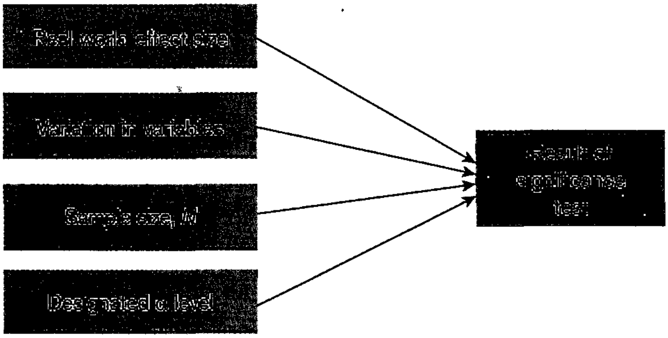
STATISTICAL SIGNIFICANCE AND THEORETICAL IMPORTANCE. We declared at the outset of the discussion on inference that people depend on the knowledge generated from samples. An integral part of that knowledge is the concept of statistical significance. Policy makers, politicians, journalists, academics, and laypeople frequently try to prove a point by claiming something to the effect that "studies indicate a statistically significant difference between A and B" or that "there is no statistically significant association between X and

Y." Hypothesis testing has become a common feature of both social and scientific discourse.

As empirical political scientists, we are happy that people resort to data and statistics to justify their positions. Nevertheless, great confusion exists about what "significance" really entails. We have given you a lot of background about what goes into hypothesis testing and the assertions that something is or is not significant. Keep in mind, though, that these tests rest on specific assumptions and procedures, and making meaningful generalizations from samples depends on how thoroughly these assumptions and procedures are satisfied.

Sometimes, when a person says that "findings are statistically significant" the implication is that a possibly earth-shattering discovery has been produced. But the hard truth is that a significance test does not prove that a meaningful effect has been uncovered. Too many other factors (as shown in figure 12-6) can cloud the interpretation of a hypothesis test.

FIGURE 12-6 **Factors That Affect Significance**



A major factor is the sample size. All other things being equal, the larger the sample, the easier it is to find significance—that is, to reject a null hypothesis. Why? The sample size does its work through the standard error, the measure of a sample estimator's precision or, loosely speaking, its closeness to the population value it estimates. If N is relatively small, sample estimates of a parameter will jump all over the place. But when N is relatively large, sample estimates tend to be close to one another. This variation shows up in the magnitude of the standard error, which in turn goes into the formulas for observed test statistics.

To demonstrate the point, suppose a null hypothesis is H_0: $\mu = 100$. Further, assume the sample mean is $\bar{Y} = 105$ and $\hat{\sigma} = 50$. (We will use the z statistic for illustrative purposes.) Let the sample size increase from 25 to 100 to 400 to 900. The observed z values are as follows:

$$z_{\text{obs1}} = \frac{(\bar{Y}-\mu)}{\frac{\hat{\sigma}}{\sqrt{N}}} = \frac{(105-100)}{\frac{50}{\sqrt{25}}} = \frac{5}{\frac{50}{5}} = \frac{5}{10} = .5.$$

$$z_{\text{obs2}} = \frac{(105-100)}{\frac{50}{\sqrt{100}}} = \frac{5}{\frac{50}{10}} = \frac{5}{5} = 1.0.$$

$$z_{\text{obs3}} = \frac{(105-100)}{\frac{50}{\sqrt{400}}} = \frac{5}{\frac{50}{20}} = \frac{5}{2.5} = 2.0.$$

$$z_{\text{obs4}} = \frac{(105-100)}{\frac{50}{\sqrt{900}}} = \frac{5}{\frac{50}{30}} = \frac{5}{1.6667} = 3.0.$$

The lesson is that a relatively small departure of a sample mean from the hypothesized mean becomes more and more statistically significant as the sample size increases, when everything else remains the same. This kind of argument backs up an old saying popular among statistics teachers: "You can *always* prove something if you take a large enough sample." A less cynical view is that "statistical significance is not the same thing as substantive significance."

Confidence Intervals and Confidence Levels

Recall that if we take many samples to obtain estimates of a population parameter, our estimates will be normally distributed and cluster around the true value of the population parameter. Sampling distributions tell us the probability that our estimates fall within certain distances of the population parameter. This probability is known as the **confidence level**. The **confidence interval** refers to the range of likely values associated with a given probability or confidence level. Thus, for every confidence level, a particular confidence interval exists.

The general form of the confidence interval is as follows:

Estimated parameter value ± standard error x critical value.

Let's return to the question of ideology in America. Look back at figure 12-3. You will see the sample mean, $\bar{Y} = 4.2696$; the standard deviation, $\hat{\sigma} = 1.4749$; and the standard error of the mean, $\hat{\sigma}_{\bar{Y}} = 1.4749/\sqrt{920} = .0486$. Let's start out with a

99 percent confidence level. Since we have a large sample, we use the table of z scores to find the critical value. We need to find the critical value associated with 1 percent ($\alpha = .01$) of the distribution in the tails. The table reports the percentage of estimates likely to fall in one tail. So we need to divide .01 by 2, which is .005. Then we look in appendix A for this value. Looking at figure 12-2, we find .049, which corresponds with a z score of 2.58. This is our critical value. Substituting 2.58 into the above equation, we find that our confidence intervals are:

$$\text{Lower} = 4.2696 - (.0486)2.58 = 4.144 \text{ and}$$
$$\text{Upper} = 4.2696 + (.0486)2.58 = 4.395.$$

Thus, we can say that we are 99 percent confident that the actual mean in the population is between the values of 4.144 and 4.395. Figure 12-3 shows the confidence interval for the 99 percent confidence level. If, for instance, you want the 95 percent interval, then α is .05. The z score defining the upper $\alpha/2 = .05/2 = .025$ tail area of the standard normal distribution is 1.96.

The same procedure is followed using the t distribution for small samples. Suppose the mean number of physicians per 100,000 people for a sample of 17 developed nations is 321 with a standard deviation of 72.4. To report the range of values that might include the unknown population mean of the number of physicians with a 95 percent confidence level, we would find the critical value for $\alpha = .05$ for a two-tailed test with 16 degrees of freedom. The critical value is 2.120. To calculate the interval, we must first calculate the standard error for the sample, which is $72.4/\sqrt{17}$ or 17.56. Thus, the 95 percent confidence interval is 321 ± 17.56 × 2.120, or between 283.77 and 385.23. You interpret these two numbers as saying, "We are 95 percent sure that the average number of physicians per 100,000 people of the developed world lies somewhere between roughly 283.77 and 358.23 physicians. We are not 100 percent positive, but the evidence is overwhelmingly in that direction." (As before, let us stress the fine point that our confidence actually lies in the method of obtaining the confidence limits—we believe that 95 times out of 100 the technique will return a pair that covers the population value.)

Here's a very important and useful tip. You can learn a lot from a confidence interval that you can't get from just one statistic: the sample mean. Confidence intervals, in a sense, suggest a span of plausible values for a parameter. By the same token, they indicate values that are implausible. For example, the interval in the previous example strongly suggests that the mean number of physicians per 100,000 persons of all the developed nations is not, say, 400; it's even less likely to be 600. Why? These values fall outside the interval. Or, to anticipate the next section, suppose someone claims that the mean for these nations is a meager 200. This argument can be considered to be a statistical hypothesis and can be tested. Since the hypothesized value falls beneath the lower confidence interval, you have reason to doubt it.

HOW IT'S DONE

The Construction of Confidence Intervals

Calculate the confidence interval for population parameter θ based on sample of size N and level of significance α:

1. Obtain an estimate of θ.
2. Find the standard error of θ, or $\hat{\sigma}_{\hat{\theta}}$.
3. Determine the critical value based on α, the desired level of significance.
4. Multiply the standard error by the critical value.
5. Add and subtract this product from the sample estimate.

Now flip around the argument. Suppose another research team estimates the average number of physicians to be 295. Although this value lies considerably below our estimate of 321, is it really inconsistent? Remember, we are pretending that the data come from random samples, so both estimates inevitably have some imprecision. Our confidence interval, however, goes from about 283 to slightly more than 358. Hence, it includes the alternative estimate. On just these grounds, we can't argue that the other estimate is wrong; depending on the sample size and variance of their sample, 321 may not fall in their confidence interval.

There is a rule of thumb in statistics, as in life: the more certain you need to be, the more information you have to have. In the case of confidence intervals, the higher the assurance you have to have, the wider your interval will be for a given sample size. Table 12-4 illustrates this point. In it we calculate 80 percent, 90 percent, 95 percent, and 99 percent confidence intervals for the mean GNP of the developing nations. The first column gives the different degrees of confidence requested. The last column shows the widths of the resulting intervals; you might loosely interpret these numbers as the "margins of error" in the estimate. Notice that as you go down the table, this error margin increases. Stated differently, if you can be content with a ballpark guess, say 80 percent intervals, then the difference between the upper and lower limits is 4,730.90 – 2,847.10 = 1,883.80, or about $1,884. But if you want to be as close as possible, or 99 percent certain, the interval becomes twice as wide, a whopping $4,076. This example is based on just nineteen cases. To tell the story one more time, the estimator of the population mean is not wrong or invalid—but it is imprecise.

If you want narrower interval widths while still being, say, 95 percent confident, then you have to increase the sample size. Table 12-5 tells you what you get for larger and larger samples. As N increases, the interval widths shrink. If you could somehow take a very large sample of developing nations—you cannot, of course, but just imagine you can—you could increase the estimator's precision from almost $3,000 to less than $500. The downside is the expense of collecting the extra data.

The bottom line is that statistical inference demands a balance between exactitude and sample sizes. If you want or need to be more exact, you need a bigger sample. But whether or not you need to be more or less exact is *not* a matter of statistics; it is essentially a substantive or practical question. Our feeling is that in an exploratory study in which the investigator is entering new territory, precision may not be as important as making sure the sample is drawn correctly and the measurements are made and recorded carefully and accurately. Only when a lot is riding on the accuracy of estimates will huge sample sizes be essential.

TABLE 12-4 **Confidence Intervals Calculated for Four Different Confidence Levels**

Sample mean = 3,789 *N* = 19			
Percent confidence	Confidence Interval		
	Upper limit	Lower limit	Interval width
80	4,730.9	2,847.1	1,883.8
90	5,017.0	2,561.0	2,456.0
95	5,276.9	2,301.1	2,975.9
99	5,827.2	1,750.8	4,076.4

TABLE 12-5 **Confidence Intervals for Various Sample Sizes at 95 Percent Level of Confidence**

Sample mean = 3,789			
Sample size (*N*)	Upper limit	Lower limit	Interval width
19	5,276.9	2,301.1	2,975.8
50	4,644.7	2,933.3	1,711.3
100	4,393.1	3,184.0	1,210.1
1,000	4,023.8	3,554.2	469.69

Conclusion

In this chapter, we started down the road to understanding statistical inference, including hypothesis testing and estimation. We leave you with some guidelines for improving your research and evaluating that of others.

No single summary statistic does or can say everything important about even a small amount of data. Consequently, you should rely on several summary measures and graphs, not just one.

Look at each variable individually. How much variation is there? What form does its distribution have? Are there any "problem" observations? Does it seem to have an association with other variables? What kinds?

If this sounds like a lot of work, think carefully *before* collecting any data. Just because a variable is in a collection doesn't mean you have to include it in your study. Ask what will be just enough to support or refute a hypothesis.

Most readers probably will not be in a position to analyze much more than six to ten variables. True, computers make number crunching easy. But it is very hard to take in and discuss in substantive terms a mass of tables, graphs, and statistics. You are probably better off studying a small dataset thoroughly than analyzing a big one perfunctorily.

Try to understand the principles of statistical inference and think continually about the topic or phenomenon being studied and the practical or real-world meaning of the results. Do not get hung up on technical jargon.

A well-thought-out and carefully investigated hypothesis that the data do not support can be just as informative and important as a statistically significant result. It is not necessary to report only "positive" findings; in fact, it's misleading. Chapter 2 discussed the roles that replication and falsification play in science. If you are studying a claim that lots of people believe and discover that your data do not support it, you will have made a positive contribution to knowledge.

TERMS INTRODUCED

Alternative hypothesis. A statement about the value or values of a population parameter. A hypothesis proposed as an alternative to the null hypothesis.

Confidence interval. The range of values into which a population parameter is likely to fall for a given level of confidence.

Confidence level. The degree of belief or probability that an estimated range of values includes or covers the population parameter.

Null hypothesis. A statement that a population parameter equals a single or specific value. Often a statement that the difference between two populations is zero.

Research or alternative hypothesis. The hypothesis in favor of which researchers usually hope to reject the null hypothesis; represented by H_A.

Statistical hypotheses. Two types of hypotheses essential to hypothesis testing: null hypotheses and research or alternative hypotheses.

Statistical significance. The probability of making a type I error.

Type I error. Error made by rejecting a null hypothesis when it is true.

Type II error. Error made by failing to reject a null hypothesis when it is not true.

***z* score.** The number of standard deviations by which a score deviates from the mean score.

SUGGESTED READINGS

Abelson, Robert P. *Statistics as Principled Argument.* Hillsdale, N.Y.: Lawrence Erlbaum, 1995.

Agresti, Alan. *An Introduction to Categorical Data Analysis.* New York: Wiley, 1996.

Agresti, Alan, and Barbara Finlay. *Statistical Methods for Social Sciences.* Upper Saddle River, N.J.: Prentice Hall, 1997.

Agresti, Alan, and Christine Franklin. *Statistics: The Art and Science of Learning from Data.* Upper Saddle River, N.J.: Prentice Hall, 2007.

Lewis-Beck, Michael. *Data Analysis: An Introduction.* Thousand Oaks, Calif.: Sage, 1995.

Velleman, Paul, and David Hoaglin. *Applications, Basics, and Computing of Exploratory Data Analysis.* Pacific Grove, Calif.: Duxbury Press, 1983.

CHAPTER 13

Investigating Relationships between Two Variables

CHAPTER OBJECTIVES

13.1 Determine how the values of one variable are related to those of another.

13.2 Explain cross-tabulation as a means of showing the relationship between categorical viable associations.

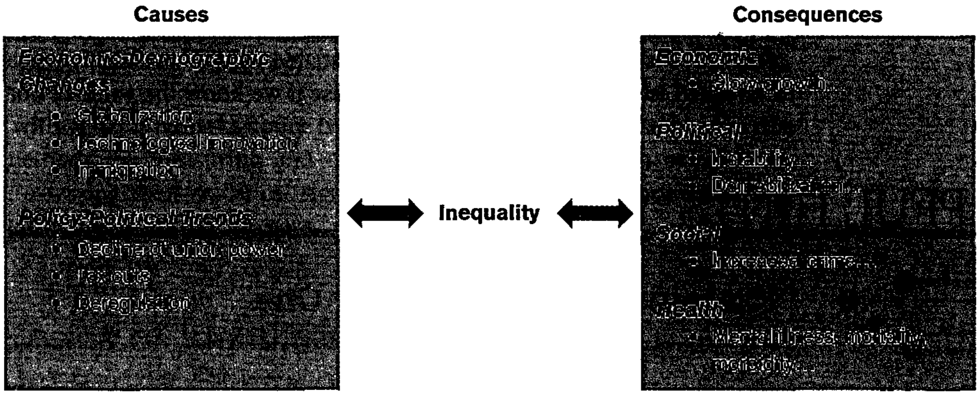
13.3 Describe the methods and tools used to measure the strength of relationships in tables.

13.4 Relate the ways to explain relationships between a categorical and a quantitative variable.

13.5 Identify the regression analysis methods for describing the ways in which an independent and dependent variable are associated.

EARLIER CHAPTERS (1, 11) raised the issue of economic inequality in the United States. Figure 13-1 shows how a political scientist might think about economic inequality. The main variable, inequality, is believed to be caused by certain factors. So, for example, one might believe that technological innovation and foreign competition have reduced low-wage jobs and hence aggravated the unequal distribution of income. Another widely discussed potential contributing factor is the decline in unions and the growth of business political power as a result of tax and regulatory reforms. The other side of the diagram shows factors that might result from growing inequality such as political instability, a loss of interest in politics, and dysfunctional social behaviors. You can interpret the arrows as saying, "There is a possible (hypothesized)

FIGURE 13-1 **How Political Scientists Think about Inequality**



Read the arrows (⟷) as "there are hypothesized relationships (causal or otherwise) between inequality and indicators of its causes and consequences."

Source: Created by authors.

relationship between each itemized indicator and inequality." As an example, Frederick Solt concludes that

> economic inequality powerfully depresses political interest, discussion of politics, and participation in elections among all but the most affluent. . . .[1]

Our task in this chapter is to show you how to investigate and verify an empirical claim like this one. Chapter 11 demonstrated ways to describe and summarize a batch of numbers, tables, descriptive statistics, and graphs to show, for example; what a "typical case" looks like (central tendency), how much variability there is in the observations (dispersion), and what the overall pattern of data looks like (shape of the distribution). And the previous chapter set up a framework for testing statistical hypotheses. We now add to this toolkit techniques for describing and measuring the association (if there is any) between two variables. Such methods are frequently called "bivariate analysis."

[1] Frederick Solt, "Economic Inequality and Democratic Political Engagement," *American Journal of Political Science* 52 (January 2008): 48.

Generally speaking, a statistical association between two variables exists if the values of the observations for one variable are connected to the values of the observations for the other. For example, if as people get older they become more politically active, then the values of the dependent variable (voting or not voting) are associated with the values of the independent variable (age). Knowing that two variables are related lets us make predictions, because if we know the value of one variable,

HELPFUL HINTS

Examine Variables One by One

Before undertaking a bivariate or multivariate analysis, examine each variable one by one. First note the types: Are they categorical (ordinal or nominal), quantitative (interval and ratio scales), or a mixture?

For categorical variables look for the following:

- Order among categories
- The modal (most frequent) category
- The distribution of cases into each category and overall shape of the distribution of cases across the categories (skewed, uni- or bimodal, etc.)
- Nearly empty categories that might be combined
- Categories not of substantive interest that can be dropped (e.g., missing value codes)

For quantitative variables look for the following:

- Missing values
- Summary statistics such as mean, median, range, or variance (standard deviation)
- Range of variables; any limits such as 0 to 1 or 0 to 100; whether negative values are possible and what they mean
- Shape of the distribution
- Substantive interpretation of scales: What does a one-unit increase or decrease in the variable mean in practical or theoretical terms?
- Any outliers or extreme values

For all variables think about the following:

- Which variables are (plausibly) dependent?
- Which variables are independent or explanatory?
- Which variables might be causal?

Use this information to pick an appropriate statistical method and interpret the results.

Check out more Helpful Hints at **edge.sagepub.com/johnson8e**

we can predict the value of the other. But many other questions arise: How much error is there in our predictions? How "strong" is the relationship? What is its direction or shape? Is it a causal one? Does it change or disappear if other variables are brought into the picture? If the relationship has been detected in a sample, can we conclude that it holds for the population?

We start with general remarks about two-variable relationships and then describe several methods for measuring and interpreting them and, when samples are involved, assessing their statistical significance. We employ both numerical and graphical techniques for this purpose.

The Basics of Identifying and Measuring Relationships

Determining how the values of one variable are related to the values of another is one of the foundations of empirical social science inquiry. This determination touches on several matters that we consider in the following sections:

- The level (or scale) of measurement of the variables: Different kinds of measurement necessitate different techniques.
- The "form" of the relationship: One can ask if changes in X move in lockstep with increases (or decreases) of Y or whether there is a more complicated connection.
- The strength of the relationship: It is possible that some levels of X will *always* be associated with certain values of Y; more commonly, though, there is only a tendency for the values to covary, and the weaker the tendency, the less the "strength" of the relationship.
- Numerical summaries of relationships: Social scientists strive to boil down all the different aspects of a relationship to a single number that reveals the type and strength of the association. These numerical summaries, however, depend on how relationships are defined.

Level of Measurement

Just as the level of measurement of a variable was important in the selection of appropriate descriptive statistics, so too is it important in selecting the appropriate method for investigating relationships between variables. Procedures for measuring relationships are summarized in table 13-1.[2]

[2] In reality, there are many techniques for analyzing relationships at a given level of measurement. The ones presented in this chapter are the most common and least complicated.

TABLE 13-1 **Levels of Measurement and Statistical Procedures: A Summary**

Type of Dependent Variable	Type of Independent Variable(s)	Procedure
Quantitative	Dichotomous*	Difference of means, boxplots
Quantitative	Categorical (nominal or ordinal) More than two	One-way analysis of variance (ANOVA) Boxplots
Categorical (nominal or ordinal)	Categorical (nominal and/or ordinal)	Cross-classification tables analysis: measures of association Log-linear models Association models
Quantitative	Quantitative and/or categorical	Linear regression Scatterplots
Dichotomous*	Quantitative and/or categorical (nominal and/or ordinal)	Logistic regression Effect plots

*A dichotomous variable has two categories.

Types of Relationships

A relationship between two variables, *Y* and *X*, can take one of several forms (use figure 13-2 as a reference).

- General association: The values of one variable—*Y*, say—tend to be associated with specific values of the other variable, *X*. This definition places no restrictions on how the values relate; the only requirement is that knowing the value of one variable helps to know or predict the value of the other. For example, if religion and party identification are associated, then certain members of certain sects should tend to identify with certain political parties. Discovering that a person is Catholic should say something about his or her partisanship. If there is no connection at all between the values of *Y* and *X*, we assert that they are *independent* of one another. (Statistical independence gets discussed in a later section of this chapter.)
- Monotonic correlation:
    - Positive: When high values of one variable are associated with high values of the other and, conversely, low values are associated with

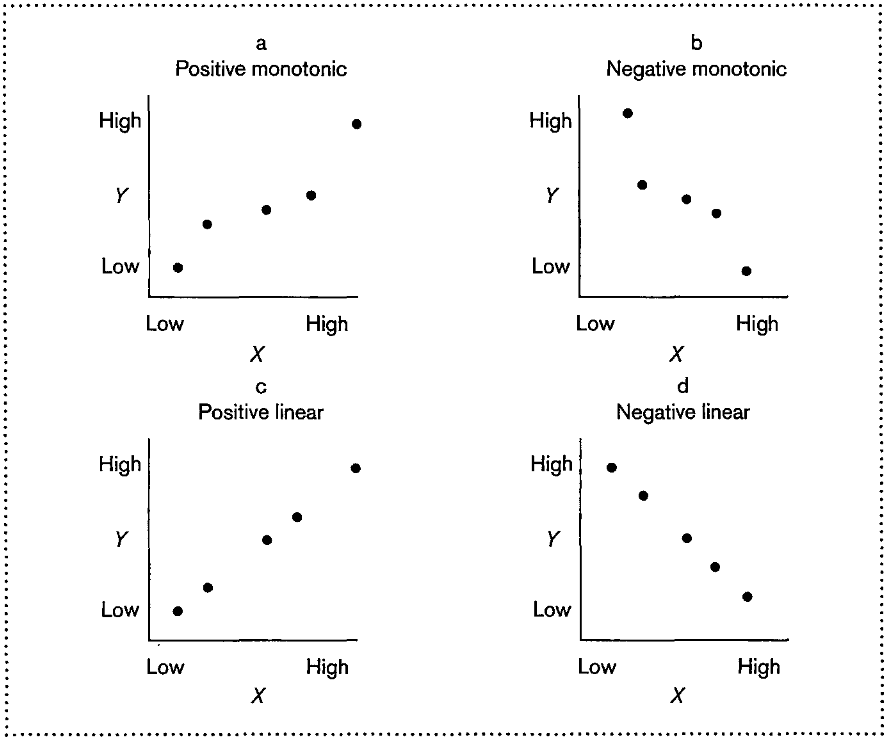
low values. On a graph, X-Y values drift upward from left to right (see figure 13-1a). A line drawn through the graph will be curved but *never goes down once it is on its way up.*

- Negative: High values of Y are associated with low values of X, and—equally—low values of Y are associated with high values of X. A graph of Y-X pairs drifts downward from left to right and *never turns back up:*

- Linear correlation: A particular type of monotonic relationship in which plotted Y-X points fall on (or at least, close to) a straight line. (Figure 13-2c shows an example of a positive linear correlation.) If the plotted values of Y and X fall on a straight line that slopes downward from left to right, the relation is called a negative correlation (see figure 13-1d).

Variables may vary together in other patterns, as when values of X and Y increase together until some threshold is met and then decline. Since these curvilinear patterns of association are hard to analyze, we set them aside in this book. In any case, the important point is that the first step in data analysis is the examination of plots to determine the approximate form or type of relationship.

FIGURE 13-2 **Types of Correlation**

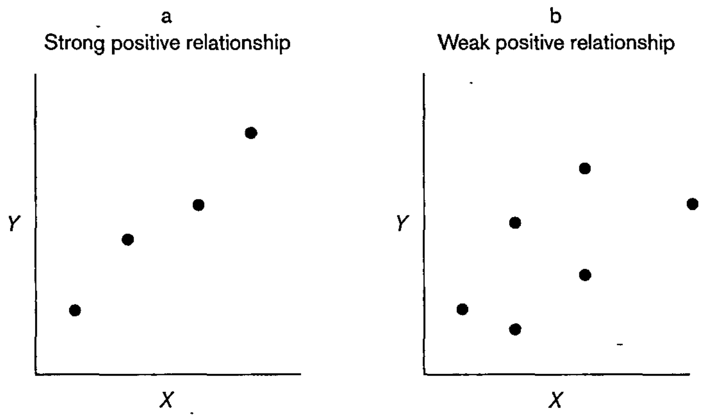


Strength of Relationship

Virtually no relationship has a cut-and-dried or "perfect" form. There are, in other words, degrees of association, so it makes sense to talk about their strength. The graph in figure 13-3 provides an intuitive idea of what is meant by "strength." Observe that in the first example (a), the values of *X* and *Y* are tied tightly together; you could even imagine a straight line passing through or very near most of the points.

In the second illustration, by contrast, the *X*-*Y* points seem spread out, and no simple line or curve would connect them. Yes, there *is* a tendency for the values to be associated—as *X* increases, so does *Y*—but the connection is rather weak.

FIGURE 13-3 **Strong and Weak Relationships**



HELPFUL HINTS

Study Graphs Carefully

We always encourage you to examine graphs of relationships among variables. But as essential as the visual devices are, they almost always need to be supplemented by numerical indices that in a "word" describe the form and strength of relationships. The general term for these statistics is *measures of association.*

Check out more Helpful Hints at **edge.sagepub.com/johnson8e**

Numerical and Graphical Procedures

As in the case of a single variable, we can employ three general types of tools to summarize and describe two-variable (or bivariate) relationships:

- Tables (e.g., two-way cross-classifications or cross-tabulations)
- Graphs (e.g., scatterplots, boxplots)
- Single numbers (e.g., measures of association, correlation coefficients)

The first two were touched on in chapter 11, so we'll come back to them later. For the moment, let's concentrate on measures of association and correlation, perhaps the two concepts one most needs to understand in order to evaluate scholarly literature and political discourse.

A **measure of association** describes in a single number or index the kind and strength of relationship between the values of two variables. The remainder of this chapter contains a half-dozen or so such indicators, and most social science programs crank them out automatically. Furthermore, to an extent that is probably misguided, these numbers are used to support theoretical or policy claims, much as the results of statistical tests are (see chapter 12). It is imperative, then, to develop a feel for what the numbers do (and do not!) say about possibly complex relationships. This requires an analyst to know the definitions of measures of association. A computer program, for example, might report that a coefficient of association between *X* and *Y* is .35. What exactly does that mean? Is there a strong or weak relationship?

At the most general level, if there is an association between, say, *X* and *Y*, then if one knows a person's particular value on *X*, it is possible to predict his or her value on *Y*. Knowing a person's gender, for example, allows a researcher to predict the individual's position on capital punishment, assuming the variables are associated. Of course, if the variables are not related according to the definition, then the coefficient will suggest that no prediction is possible. The coefficients we describe in this chapter (1) assume a particular level of measurement—nominal, ordinal, interval, or ratio, and (2) rest on a specific conception of association. Stated differently, each coefficient measures a specific type of association, and to interpret (translate) its numerical value into everyday language, you have to grasp the kind of association it is measuring. Two variables can be strongly associated according to one coefficient and weakly (or not all) by another. Therefore, whenever we describe a measure such as the correlation coefficient, we need to explain what kind of relationship it is intended to measure.

Here are some important properties of commonly used coefficients:

- Null value: Usually, but not always, zero indicates no association, but there are important exceptions to this rule of thumb.

- Maximum value: Some coefficients do not have a maximum value; they can be in theory very large. Many, however, are bounded: normally their upper and lower limits are 1.0 and –1.0. When a coefficient attains a bound, variables are said to be perfectly associated according to the coefficient's definition.
- Strength of relationship: Subject to lower and upper boundaries, a coefficient's absolute numerical value increases with the strength of the association. So, for example, a coefficient of .6 would indicate a stronger relationship than one of .3. (But the relationship would not necessarily be twice as strong. It all depends on how the statistic is defined.)
- Level of measurement: As indicated above, nominal, ordinal, and quantitative (ratio and interval) variables each require their own type of coefficient. You can, of course, pretend that ordinal scales are numeric and calculate a statistic intended for quantitative data—plenty of people do—but since lots of research has gone into measures of association for different levels of measurement and satisfactory alternatives exist, you should be able to find one or two that will fit your data.
- Symmetry: The numerical magnitudes of some indices depends on which variable, *Y* or *X*, is considered independent. These are asymmetric measures. The value of a coefficient calculated with *Y* as dependent may very well differ from the same indicator using *X* as the dependent variable. A symmetric measure keeps the same value no matter which variable is treated as dependent or independent.
- Standardized or unstandardized: The measurement scale on which variables are measured affects the numerical value of some measures, whereas others are not so affected.

Table Summaries of Categorical Variable Associations

A **cross-tabulation** shows the joint or bivariate relationship between two categorized (nominal and/or ordinal) variables. Here we are not dealing with a single number but rather with an array of frequencies, or proportions, or percentages. With categorical data such a tabulation is usually more interesting than a single index because one sees how specific categories of one variable are tied to those of the other. Later, we present single-number coefficients, but they are best used with a table.

To start the explanation, let's return to a previous topic, political participation. Studies cited in the first chapter have found that several variables—socioeconomic status, political interest, and partisanship among them—affect the decision to vote or stay at home.

An obvious hypothesis is "The greater one's party loyalty, the greater one's willingness to spend time and money on politics." We can use the partisanship scale developed in chapter 11 to demonstrate how a cross-tabulation can assist in hypothesis testing. Recall that this indicator attempts to tap into the degree or intensity of partisan feelings, not their political direction. Hence, "independents" are coded 1; "leaning Democrats or Republicans" get 2; those who simply identify with either party, but not strongly, receive scores of 3; and finally, the strong partisans of either party (those who said they "strongly" identified with their party) are 4. This is an ordinal scale that extends from 1, "least partisan," to 4, "most partisan," with (we hope) more or less equally spaced psychological levels in between. For the moment, we are not going to take advantage of the quasinumerical scale and instead treat it as a simple categorical variable.

The data in table 13-2 provide a simple test of the hypothesis that partisanship is related to political activity—in this case, donating money to a political organization or cause. "Reading" the table is straightforward. The sample consists of 134 nonpartisans, those who state no preference for either party. Similarly, there are 141 "weak" partisans who lean toward one or the other party but do not identify with either; 359 moderates (Democrats or Republicans); and 322 highly partisan individuals, those who use the adjective *strong* to describe their party affiliations.

TABLE 13-2 **Level of Partisanship: Donating Money**

Donated Money?	Nonpartisan (Independent)	Weak	Moderate	Strong
No, did not donate	91.8% (123)	71.6% (101)	87.5% (314)	68.0% (219)
Yes, donated	8.2% (11)	28.4% (40)	12.5% (45)	32.0% (103)
Totals	100% (134)	100% (141)	100% (359)	100% (322)

Question: "During the last 12 months, have you done any of the following? Donated money to a political organization or group."

Note: Cell entries are percentages and (frequencies).

Source: United States Citizenship, Involvement, Democracy Survey, 2006.

(See the row labeled "Totals.") Next consider how the independents are distributed on the dependent variable, donated or not. We see that 123, or about 92 percent (123/134 ≈.92), of them report not having contributed in the last year. By the same token, fewer than 9 percent did donate. (Look at the next row down marked "Yes, donated.") Going across the "Yes" row, it is apparent that as partisanship increases, so too does the proportion of respondents who contribute. In fact, we find a 24 percentage point difference between strong and nonpartisan: 8 percent versus 32 percent. The behavior of those in the middle partisan categories fits the pattern, except that for some reason the "weak" group gives more than the "moderates" do. (These sorts of anomalies arise all the time in survey research and invite us to think carefully about our measurements and data analysis. Did we make a coding mistake, for instance?) Overall, then, we could conclude that the research hypothesis is tenable. We might even say there is a positive monotonic correlation (with the exception just noted).

HOW IT'S DONE

Building a Cross-Tabulation

Suppose you have a dependent or response variable (*Y*) with two categories, A and B, and another (*X*) with three levels, L, M, and H. To construct a cross-tabulation, find each observation's scores on *Y* and *X* and the cell in a table that corresponds to these values. Mark the observation's position with a tally mark ("/"). Do the same for all *N* cases and write the totals in each cell. Add across the rows and then down the columns to obtain the row and column *marginal* totals. Convert to proportions or percentages as needed. Make sure each combination of *Y*-*X* scores appears in one and only one cell. Of course, hardly anyone follows this "by-hand" procedure except the smallest polls. The work is done with electronic processing equipment. But this is essentially the logic computers follow. And knowing the underlying process may further clarify the table's meaning.

	Categories of variable X			
Categories of variable Y	**L**	**M**	**H**	**Totals of row tallies**
A	*𝍸 //* 7	*𝍸 𝍸 //* 12	*𝍸* 5	24
B	*//* 2	*𝍸 /* 6	*𝍸 ///* 8	16
Totals of column tallies	9	18	13	40

HELPFUL HINTS

Categories with "Too Few" Cases

A widely accepted rule of thumb asserts that percentages based on twenty or fewer observations are not reliable indicators and should not be reported or should be reported with "warning signs." Suppose, for example, a survey contained only fifteen respondents in a category of the independent variable, such as Asian Americans. If you try to find the percentage of this group that identify as, say, strong Republicans, the resulting estimate will be based on such a small number (15) that many readers and analysts may not have confidence in it. Two possible solutions come to mind. First, use a symbol (for example, †) to indicate "too few cases." Alternatively, the category could be combined with another one to increase the total frequency. The text gives some examples.

Check out more Helpful Hints at **edge.sagepub.com/johnson8e**

The previous example was especially simple because we skipped some nuances such as what to do with missing values. Cross-tabulations are a foundation of data analysis, and grasping what they reveal (and do not reveal) about relationships will stand you in good stead as a student of political science and politics.

Here is a slightly more complicated example. In the wake of the 2010 midterm elections, there was a lot of talk about growing polarization in American politics. It might be worthwhile to investigate the behavior and attitudes of highly partisan voters vis-á-vis their less engaged neighbors. The Citizenship, Involvement, Democracy (CID) survey used in table 13-3 asked, "Which statement best describes your preference: Politics should be about finding a compromise between people with different views OR Politics should be about sticking to your convictions, and fighting to implement them." Given the purported increase in divisiveness in politics, one might hypothesize "The greater the feelings of party loyalty, the less the willingness to compromise one's principles."[3] Table 13-3 provides a tentative answer.

[3] Chapter 11 discusses the construction of this variable. Briefly, measuring partisanship that is based on responses to a question asking respondents if they identified with a party, responses coded "strong" Democrats and Republicans were classified as most partisan, independents least partisan. Those with weak to moderate party attachments fell in the middle two categories.

TABLE 13-3 **Level of Partisanship: Finding Compromise versus Sticking to Your Convictions**

Position on Compromise	Nonpartisan (Independent)	Weak	Moderate	Strong
Finding compromise	40.91% (54)	40.43% (57)	32.50% (117)	31.15% (100)
Lean toward finding compromise	26.52 (35)	35.46 (50)	34.72 (125)	33.64 (108)
Can't say	5.30 (7)	4.96 (7)	5.28 (19)	5.92 (19)
Lean toward sticking to convictions	9.85 (13)	9.22 (13)	11.94 (43)	15.58 (50)
Sticking to convictions	17.42 (23)	9.93 (14)	15.56 (56)	13.71 (44)
Totals	100% (132)	100% (141)	100% (360)	100% (321)

Question: Respondents were asked to respond to this statement: "Politics should be about finding a compromise between people with different views OR Politics should be about sticking to your convictions, and fighting to implement them."

Source: United States Citizenship, Involvement, Democracy Survey, 2006.

Note: Cell entries are percentages and (frequencies).

Analyze the table as before by thinking about what the percentages refer to. In the row titled "Finding compromise," we see that about 30 to 40 percent of *all* partisan groups choose to compromise rather than dig in. Now go to the next row, to the category "Lean toward finding compromise." Again, the percentages are not too different. Indeed, combining the first two rows reveals that the overwhelming majority of all partisanship levels support trying to compromise. The strongest partisans (those in the last column) seem to think as everyone else does.

This table, incidentally, is an example of a "negative" finding. Such *nil* relationships often go unreported because they seem not have discovered anything important. Before burying such findings, however, an analyst should think carefully about two things:

1. Was the hypothesis stated accurately? Do the operational indicators—the questions on the survey—adequately capture the meaning of the concepts? Were the data correctly coded and analyzed? In other words, might some deficiency in the research design or analysis have led to the false rejection of the proposition?

2. As important, if the hypothesis really is viable and correctly tested, is the commonplace belief about growing hostility in American politics exaggerated? Perhaps the "pundit" class—those who are interested and active in politics—divide more sharply on party lines than the public does. If so, polarization may be an "elite," not mass, phenomenon.

Whatever the case, don't give up on a hypothesis because data do not support it. Sure, it may have been a rotten idea to begin with. On the other hand, there may be substantively and methodologically interesting and important reasons for the nonassociation.

When the categories of the independent variable are arrayed across the *top* of the table—that is, they are the column labels—it is essential that the percentages add to 100 down the columns. These are called *column percentages*. You might think of the respondents in each column as a subsample. Look at table 13-4, which compares male and female party affiliation. Suppose we want to know how the males differ among themselves on partisanship. It is necessary to use the ćolumn totals as the bases (denominators) for the percentage calculations. Thus, for the 581 men, the percentage identifying as "strong Democrats" (11.9%) *plus* the percentage identifying as "weak Democrats" (15.0%) *plus* the percentage identify as "leaning Democrat" (18.6%) . . . and so forth down through all the response categories equals 100 percent. The same is true for women: the total of column percentages sums to 100 percent. It is this arrangement of percentages that allows us to compare the relative frequencies of responses between men and women.

TABLE 13-4 **Cross-Tabulation of Gender by Party Identification**

	Gender	
Party Identification Response category	**Male**	**Female**
Strong Democrat	11.9% (69)	20.9% (128)
Weak Democrat	15.0% (87)	16.2% (99)
Independent-leaning Democrat	18.6% (108)	16.5% (101)
Independent	10.2% (59)	9.3% (57)
Independent-leaning Republican	14.5% (84)	9.1% (56)
Weak Republican	13.9% (81)	11.1% (68)
Strong Republican	16.0% (93)	17.0% (104)
Total *N*=1,194	100.1% (581)	100.1% (613)

Source: 2004 National Election Study.

Note: Totals subject to rounding error.

Suppose you asked a computer to give you percentages by row totals, i.e., *row percentages*. Table 13-5 suggests what might result, and the possible difficulties of interpretation. If you were not careful, you might conclude that there was a huge gender difference on "strong Democrat," 35 percent versus 65 percent. But this is not what the numbers mean. There are 197 strong Democrats in the sample (look in the last column), of which 35 percent are men and 65 percent women. It would be reasonable to say that strong Democrats tend to be women, whereas independents are about half male and half female. Still, if in your mind one variable (e.g., party identification) depends on another variable (e.g., gender) and you want to measure the effect of the latter on the former, make sure the percentages are based on the independent variable category totals.

TABLE 13-5 **Row Percentages Are Not the Same as Column Percentages**

Party Identification Response Category	Gender		
	Male	**Female**	**Total**
Strong Democrat	35.0% (69)	65.0% (128)	100% (197)
Weak Democrat	46.8% (87)	53.2% (99)	100% (186)
Independent-leaning Democrat	51.7% (108)	48.3% (101)	100% (209)
Independent	50.9% (59)	49.1% (57)	100% (116)
Independent-leaning Republican	60.0% (84)	40.0% (56)	100% (140)
Weak Republican	54.4% (81)	45.6% (68)	100% (149)
Strong Republican	47.2% (93)	52.8% (104)	100% (197)

Source: 2004 National Election Study.

Note: Numbers in parentheses are frequencies.

Measuring Strength of Relationships in Tables

Do the data in table 13-4 support the hypothesis of a "gender gap"? As we just indicated, a careful examination of the column percentages suggests that the hypothesis has only minimal support. Why? Because a scrutiny of the partisanship distributions by gender does not show much difference. Yet it would be desirable to have a more succinct summary, one that would reveal the strength of the relationship between gender and party identification.

The strength of an association refers to how different the observed values of the dependent variable are in the categories of the independent variable. In the case of cross-classified variables, the strongest relationship possible between two variables is one in which the value of the dependent variable for every case in one category of the independent variable differs from that of every case in another category of the independent variable. We might call such a connection a *perfect relationship,* because the dependent variable is perfectly associated with the independent variable; that is, there are no exceptions to the pattern. If the results can be applied to future observations, a perfect relationship between the independent and dependent variables enables a researcher to predict accurately a case's value on the dependent variable given a known value of *X*.

A weak relationship would be one in which the differences in the observed values of the dependent variable for different categories of the independent variable are slight. In fact, the weakest observed relationship is one in which the distribution is identical for all categories of the independent variable—in other words, one in which no relationship appears to exist.

To get a better handle on strong versus weak relationships as measured by a cross-tabulation, consider the hypothetical data in tables 13-6 and 13-7. Assume

TABLE 13-6 **Example of a Nil Relationship between Region and Opinions about Comprehensive Immigration Reform**

	Region			
Opinion	**East**	**Midwest**	**South**	**West**
Favor immigration reform	48%	48%	48%	48%
Do not favor immigration reform	52%	52%	52%	52%
Total	100%	100%	100%	100%

Note: Hypothetical responses to the question, "Do you favor comprehensive immigration reform?"

we want to know if a connection exists between people's region of residency and attitudes about immigration. (The hypothesis might be that southerners and westerners are less favorable than citizens in other parts of the country.) The frequencies and percentages in table 13-6 show no relationship between the independent and dependent variables. The relative frequencies (that is, percentages) are identical across all categories of the independent variable. Another way of thinking about nil relationships is to consider that knowledge of someone's value on the independent variable does not help predict his or her score on the dependent variable. In table 13-6, 48 percent of the easterners "favor reform," but so do 48 percent of the westerners, and for that matter, so do 48 percent of the inhabitants of the other regions. The conclusions are that (1) slightly more than half of the respondents in the survey want changes in immigration laws, and (2) there is *no* difference among the regions on this point. Consequently, the hypothesis that region affects opinions would not be supported by this evidence.

Now look at table 13-7, in which there is a strong—one might say nearly perfect—relationship between region and opinion. Notice, for instance, that 100 percent of the easterners and Midwesterners favor comprehensive change, whereas 100 percent of the southerners and westerners do not.

Most observed contingency tables, like table 13-5, fall between these extremes. That is, there may be a slight (but not nil) relationship, a strong (but not perfect) relationship, or a "moderate" relationship between two variables. Deciding which is the case requires the analyst to examine carefully the relative frequencies and determine if there is a substantively important pattern. When asked, "Is there a relationship between *X* and *Y*?" the answer will usually not be an unequivocal yes or no. Instead, the reply rests on judgment. If you think yes is right, then make the

TABLE 13-7 Example of a Perfect Relationship between Region and Comprehensive Immigration Reform

	Region			
Opinion	East	Midwest	South	West
Favor immigration reform	100%	100%	0%	0%
Do not favor immigration reform	0%	0%	100%	100%
Total	100%	100%	100%	100%

Note: Hypothetical responses to the question, "Do you favor comprehensive immigration reform?"

case by describing differences among percentages between categories of the independent variable. If, however, your answer is no, then explain why you think any observed differences are more or less trivial. A little later in the chapter, we present some additional methods and tools that help measure the strength of relationships.

Direction of a Relationship

In addition to assessing the strength of a relationship, one can also examine its "direction." The **direction of a relationship** shows which values of the independent variable are associated with which values of the dependent variable. This is an especially important consideration when the variables are ordinal or have ordered categories such as "high," "medium," and "low" or "strongly agree" to "strongly disagree," or the categories can reasonably be interpreted as having an underlying categorical spectrum, such as "least" to "most" liberal.

Table 13-8 displays the relationship between a scale of political liberalism (call it *X*) and a measure of opinions about gun control (*Y*). Both variables have an inherent order. The ideology variable can be thought of as running from lowest to highest liberalism, while responses to the question about firearms might be considered as going from least to most restrictive control.[4]

Take a moment to study the numbers in the table; we guarantee it will pay off in the long run. Start with the "most" liberal category. About two-thirds of respondents in this category (65.4%) are also "most" supportive of restricting gun purchases.

[4] These labels represent an interpretation we have imposed on the question responses. It would be perfectly legitimate, for instance, to redefine the ideology scale as the "degree of conservatism." What matters is that you keep straight in your mind how the variables are treated and make your explanations consistent with that definition.

TABLE 13-8 **Attitudes toward Gun Control by Liberalism**

	Liberalism Scale (*X*)			
Make it easier or harder to buy a gun (*Y*)	**Least conservative**	**Medium (middle of the road)**	**Most conservative**	**Total**
Least favorable to guns (make it much harder to buy)	65.5% (72)	43.5% (226)	28.2% (50)	43.2% (348)
Medium (make it harder)	14.5% (16)	17.0% (88)	7.9% (14)	14.6% (118)
Most favorable to guns (make it easier to buy plus "same as now")	20.0% (22)	39.5% (205)	63.8% (113)	42.2% (340)
Total	100% (110)	100% (519)	100% (177)	100% (806)

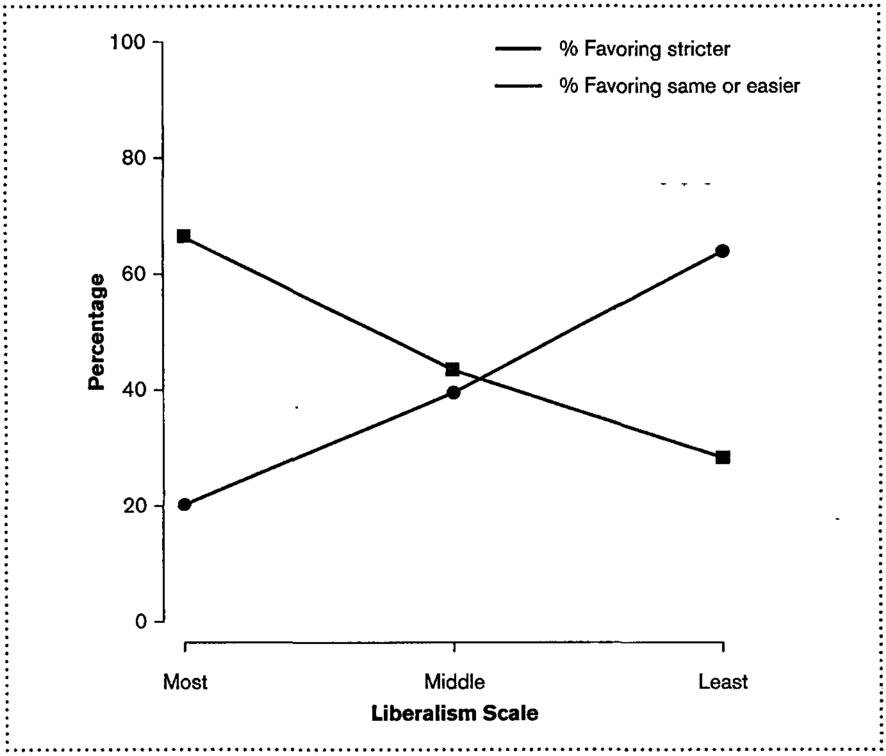
Source: 2004 National Election Study.

That is, there is a tendency for "high" values of ideology to be associated with a "high value" of gun control. Now look in the last column, the "most conservative." You should see that a clear majority of these respondents (63.8%) are in the "least" enthusiastic category of Y, the dependent variable. Here, we have a case of "low" values tending to be linked to "low" values. The middle group (independent thinkers, maybe) are more or less split between being for and against making it more difficult for people to buy firearms.

Sometimes it helps to draw a sketch of the results. Consider the top row. The percentages decline as one moves from "least" (65.5%) to "most" (28.2%) conservative. If you plot these numbers on a simple *X*-*Y* graph with equally spaced intervals for the *X* variable, you can see that the line decreases almost linearly, which can be interpreted simply as "The more conservative a person, the less favorable he or she feels toward stricter gun laws." (The percentage of each category saying "stricter" declines precipitously as one moves from liberals to independents to conservatives.) The upward-sloping line (positive slope) can be interpreted similarly. It shows the percentages in the third row, "make laws easier or keep the same," are plotted on the line that slants upward from left to right, which can be read as "The more conservative (the less liberal) an individual, the less favorable to controls" (see figure 13-4). In both instances, we see at least a monotonic correlation. (If you were to plot the middle-row percentages, what would the line look like on the graph?)

We should add that the association between these two variables, although not perfect by the standards set forth earlier, is quite strong. Why this conclusion? As a preview of things to come, try this thought experiment. Suppose you were asked to predict how Americans would respond to a question about making gun control tougher. In the absence of any other information, you might take the "marginal" distribution of responses to the question in table 13-8 as a first approximation. (The marginal totals are in the rightmost column of the table.) Thus, you could reply, "Well, most citizens are either for stricter controls (43.2%) or for leaving things as they are (42.2%), with a smattering of people (14.6%) in between." But suppose that you *also* knew people's political inclinations. This knowledge would help you improve your predictions, because the least conservative (most liberal) individuals are apt to want stronger controls, while conversely the most conservative (least liberal) respondents by and large favor leaving matters as they stand. So knowing a person's ideology enhances your predictive power. This idea—the proportional reduction in error—underlies several measures of association we will discuss shortly.

FIGURE 13-4 **Simple Interpretation of Table Percentages: Liberalism and Gun Control**



Source: Table 13-8.

HOW IT'S DONE

Computing Ordinal Measures of Association

Let C = number of concordant pairs,

D = the number of discordant pairs,

T_X = the number of pairs tied only on X,

T_Y = the number of pairs tied only on Y,

T_{XY} = the number of pairs tied on both X and Y, and

m = the minimum of I or J, where I and J are the numbers of categories of Y and X, respectively.

Gamma: $\hat{\Upsilon} = \dfrac{(C-D)}{(C+D)}$

Tau-*b*: $\hat{\tau}_b = \dfrac{(C-D)}{\sqrt{(C+D+T_Y)}\ \sqrt{(C+D+T_X)}}$

Tau-*c*: $\hat{\tau}_c = \dfrac{(C-D)}{N^2\left[\dfrac{(m-1)}{2m}\right]}$

Somers' *D*: $D_{YX} = \dfrac{(C-D)}{C+D+T_Y}$

Somers' *D*: $D_{XY} = \dfrac{(C-D)}{C+D+T_X}$

Assessing both the strength and type (direction) of a relationship in cross-classification tables requires looking at relative frequencies (percentages) cell by cell. That is not at all a bad practice. But statisticians have developed sophisticated methods for distilling the frequencies down to single numbers or "modeling" them in such a way that hard-to-see features become apparent. We next introduce a few of the ideas.

Coefficients for Ordinal Variables

So far we have examined the relationship between two categorical variables by inspecting percentages in the categories of the independent variable. To fathom their messages, we have used rough sketches and visual inspection of the tables themselves. However, if the analysis involves many tables or tables that have many cells, another way of summarizing the information is needed. Here we introduce four correlation coefficients for ordinal variables.

These statistics, much like the descriptive statistics given in chapter 11, represent the data in a table with a single summary number that measures the strength and direction of an association. (You might want to review the introductory section that lists the properties of these indicators.) Among the most common statistics are **Kendall's tau-*b*, Kendall's tau-*c*, Somers' *D*** (two versions), and **Goodman and Kruskal's gamma**—named after the individuals who developed them. Most computer programs calculate these and other coefficients as well. They are similar, but not identical, in how they summarize the contents of a two-way frequency table.

TABLE 13-9 **Table with Concordant, Discordant, and Tied Pairs**

	Variable *X*		
Variable *Y*	**High**	**Medium**	**Low**
High	Alex	Dawn	Gus
Medium		Ernesto	Hera
Low	Carl	Fay	Ike Jasmine

We will not go into the details of their calculation, partly because software makes them so readily available, but instead concentrate on their numerical meaning. Nevertheless, a bit of background won't hurt. Each coefficient compares *pairs* of cases by determining whether those pairs are "concordant," "discordant," or "tied." These can be slippery concepts, so look at table 13-9. It contains nine individuals (cases).

- A *concordant pair* is a pair in which one individual is *higher* on *both* variables than the other case. Alex and Ernesto are concordant because Alex is higher on *Y* and *X*. Alex is also concordant with Fay, Hera, Ike, and Jasmine. There are other concordant pairs such as Dawn-Hera and Ernesto-Ike.
- A *discordant pair* is one in which one case is *lower* on one of the variables but *higher* on the other. Gus, for example, has a higher score on *Y* but a lower score on *X* compared to either Ernesto, Fay, or Carl. Therefore, these pairs "violate" the expectation that as one variable increases, so does the other.
- A *tied pair* is a pair in which both observations have the same value on one or both variables. There are lots of tied pairs in this table: Alex and Dawn are tied on *Y* (they both are in the "high" category"), Alex and Carl are tied on *X* (but not *Y*), and Ike and Jasmine are tied on both *X* and *Y*. (There are several others in the table.).

All of the ordinal coefficients of association (tau-*b*, tau-*c*, Somers' *D*, and gamma) use the probability or number of pairs of different kinds to summarize the relationship in a table. In a population, they measure the probability of a randomly drawn pair of observations being concordant minus the probability of being discordant with respect to *Y* and *X*:

$$\text{Measure} = p_{\text{concordance}} - p_{\text{discordance}},$$

where *p* means probability. They differ only in whether the probabilities are conditional on the presence or absence of ties. Gamma, for example, is defined as

$$\gamma = p_{\text{C}|\text{noties}} - p_{\text{D}|\text{noties}}.$$

In plain language, it is the probability that a randomly drawn pair will be concordant on Y and X, given that it is not tied, minus the corresponding probability of discordance. An "excess" of concordant pairs over discordant pairs suggests a positive relationship; if discordant pairs are more likely, then the correlation will be negative.

In samples, the basic comparison made is between the number of concordant and discordant pairs. If both types of pairs are equally numerous, the statistic will be zero, indicating no relationship. If concordant pairs are more numerous, the coefficient will be positive; if discordant pairs outnumber concordant pairs, the statistic will be negative. The degree to which concordant or discordant pairs predominate, or one kind of pair is more frequent than the other, affects the magnitude of the statistic. Hence, if only the main diagonal were filled with observations, all the pairs would be concordant, and the statistic would be +1—a perfect, **positive relationship** (see table 13-10a). If only the minor (opposite) diagonal were filled with observations, all the pairs would be discordant, and the statistic would be –1—a perfect, **negative relationship** (see table 13-10b).

Gamma can attain its maximum (1 or –1) even if not all of the observations are on the main diagonal because it ignores all tied pairs. The other measures (tau-*b,* for example) "discount the strength of the relationship by the number of ties in the table."[5] Hence, in table 13-11, gamma would be 1.0, whereas the other coefficients would be slightly less.

In a "real" contingency table, there will be many pairs of all sorts, and counting them can be a nuisance. So we leave their computation to the computer. The formulas for these measures have the same form: one quantity divided by another. The numerator is always the number of concordant minus discordant pairs ($C - D$). The denominators differ, however, in how they handle ties. Gamma ignores tied pairs altogether, whereas the others incorporate them in different ways.[6] To help you understand them, we list a few of their properties.

- Theoretically, all vary between –1 and 1, with 1 indicating a perfect positive (monotonic) correlation and –1 a perfect negative (monotonic) correlation.
- In practice, you will most likely never see one of these coefficients attain these bounds. Indeed, even for strongly related variables, the numerical values will usually be far from 1 or –1. If any of them reaches, say, .4 or .5 in absolute value, there is an association worth investigating.

[5] You might think of ties as a "penalty" for the imprecise measurement classification involves. But however they are interpreted, tied pairs count against all the measures except gamma in the sense that the more ties, the smaller the numerical value of the coefficient. See H. T. Reynolds, *The Analysis of Cross-Classifications* (New York: The Free Press, 1977): 69–79.

[6] For further information about the calculation of each of these statistics, see Alan Agresti and Barbara Finlay, *Statistical Methods for the Social Sciences,* 3rd ed. (Upper Saddle River, N.J.: Prentice Hall, 1997), 272–82.

TABLE 13-10 **Perfect Positive and Negative Relationships**

	a. Every pair concordant (perfect positive relationship)		
	Variable *X*		
Variable *Y*	**High**	**Medium**	**Low**
High	Arthur		
Medium		Candy	
Low			Ed

	b. Every pair discordant (perfect negative relationship)		
	Variable *X*		
Variable *Y*	**High**	**Medium**	**Low**
High			Faith
Medium		Guy	
Low	Hilary		

- Since zero means no correlation, values in the range of −.1 to .1 suggest a weak relationship.
- All will have the same sign.
- The absolute value of gamma ($\hat{\Upsilon}$) will always be greater than or equal to that of any of the others. The relationships among tau-*b*, tau-*c*, and Somers' *D* are harder to generalize because they are affected differently by the cross-classification's structure (i.e., number of rows and columns).
- Somers' *D* is an "asymmetric" measure because its value depends on which variable is considered dependent. Therefore, there are really two possible versions: one, D_{YX}, has *Y* as the dependent variable, while the other,.D_{XY}, treats *X* as dependent.
- By themselves, the measures are not sufficient to assess how and how strongly one variable is related to another. You should ask the software to calculate all the coefficients *and* spend time visually inspecting the relative frequencies in the table.[7]

[7] Partly because these coefficients do not generally describe the complexities of relationships between categorical variables, they have fallen out of favor with many social scientists. Sociologists and statisticians have developed methods for modeling the multiplicity of interactions often found among categories in a table. We touch on a few techniques later in the chapter but leave the bulk of them to more advanced texts. A good introduction is Alan Agresti, *Analysis of Ordinal Categorical Data* (New York: Wiley, 1984).

TABLE 13-11 **Perfect Monotonic Relationship**

	Variable *X*			
Variable *Y*	**Very high**	**Medium high**	**Medium low**	**Very low**
Very high	Abe			
Medium high		Bertha		
Medium low			Claudio	
Very low				Darby

Gamma ($\bar{\gamma}$) = 1.0.

The last point is worth emphasizing. None of the coefficients is appropriate if the relationship "curves," in the sense that as *X* increases so does *Y* up to a certain point when an increase in *X* is accompanied by a decrease in *Y.* Consider table 13-12, which contains four observations. There is a "perfect" association: you tell me a person's value on *X*, and I will predict exactly her score on *Y.* Yet the number of concordant pairs (3) equals the number of discordant ones (3), so their difference is zero. This difference (*C* – *D*) appears in the numerator of all the coefficients, so they would all be zero, implying no relationship. But there is an association; it's just not a correlation.

TWO EXAMPLES. Hypothetical data help establish the basic ideas of these ordinal measures of association, but when push comes to shove they do not give much practice understanding actual survey results. Therefore we provide two more tables that explore questions touched on earlier. The first is a cross-tabulation of

TABLE 13-12 **Perfect but Not Monotonic Relationship**

	Variable *X*			
Variable *Y*	**Very high**	**Medium high**	**Medium low**	**Very low**
Very high				Doris
Medium high	Adele			
Medium low		Barbara		
Very low			Connie	

TABLE 13-13 **2008 Presidential Vote by Party**

	Political Ideology						
	1 Least conservative	2	3	4 Middle	5	6	7 Most conservative
Obama	98.80% (56)	94.83% (167)	80.58% (118)	61.44% (203)	30.70% (60)	10.89% (34)	10.17% (5)
McCain	1.20 (1)	5.17 (9)	19.42 (28)	38.56 (127)	69.30 (134)	89.11 (275)	89.83 (46)
Totals	100% (57)	100% (176)	100% (146)	100% (330)	100% (194)	100% (309)	100% (51)

Question: "Where would you place yourself on this (liberalism-conservatism) scale, or haven't you thought much about this?"

Chi square = 517.99; 6 *df*; gamma = 0.818; tau-*b* = .564; Somers' D_{yx} = .719, $\hat{\lambda}$ = .575.

Source: The American National Election Studies (ANES; www.electionstudies.org). The ANES 2008 Time Series Study, Stanford University and the University of Michigan (producers).

voting in the 2008 presidential election by self-placement on a seven-point liberalism-conservatism scale. The voting variable has only two categories (Barack Obama, Democrat, and John McCain, Republican), but any dichotomous variable (a variable with two categories) can be considered ordinal. You can construe the other variable as measuring the "degree" of conservatism. Since there are 7 x 2 = 14 relative frequencies to scrutinize, measures of (monotonic) correlation may help us decide how closely ideology predicts candidate preference. This table (table 13-13) is interpreted exactly like all the others: compare categories of ideology by the percentage in each who voted for, say, Obama.[8]

You should be able to detect a clear-cut pattern: as conservatism increases across the table, the propensity to vote for McCain also increases. Examine the percentages. (Notice, by the way, that the "least" and "most" conservative categories have relatively few cases in them. We might have combined those cases with the adjacent categories to improve the precision or reliability of the cell proportion estimates.)

All the measures are "large" by the standards of categorical data analysis. Gamma is 0.82, which indicates a strong positive correlation. (Why positive?) Consider the two variables as having an order: ideology runs from low to high conservatism. It is also legitimate to think of vote as having a numerical dimension, with Obama

[8] By the way, these and the other survey data have been "weighted" in order to ensure that the final samples approximate the US population. Chapter 7 explains the rationale for weighted samples. The only wrinkle is that sometimes we report fractional frequencies. Just round these to the nearest whole number.

arbitrarily used as a low value and McCain as high. Consequently, moving along the columns from left to right, we see "low" values of conservatism associated with "low" values of vote (Obama) and high conservatism scores associated with "high" on voting (McCain). It may seem strange, but a dichotomous or two-category variable can often be interpreted this way. As we mentioned earlier, these numbers seldom get close to their maximums (|1.0|), and values over .4 to .5 indicate a strong correlation. So taken together, these suggest that ideology is highly correlated with voting. Overall, the conclusion is that position on the liberalism-conservatism spectrum predicts voting. Note, however, that since the data show only covariance and not time order or the operation of other variables, we cannot say this is a causal connection.

To wrap up this section, let us look at the second example, which returns to the idea of a gender gap: Are women more liberal than men, and if so, on what issues? Here the response variable is attitudes toward allowing gays to serve in the military. (These data too come from the 2008 ANES study used earlier.) Table 13-14 shows how gender relates to preferences about gays serving in the military. Conventional wisdom might say that women will be somewhat more open to the idea than men will.

The pattern here might be a bit harder to detect. Step back for a second and look at the column totals, as usual. In raw frequencies, there are more women in the sample than men, a common result in public opinion research. Still, there are enough of each gender to make meaningful comparisons. Note first of all that the vast majority of these respondents (55% + 23% = 78%) favor strongly or simply favor allowing gays to enlist in the military (the last column contains these totals). So right away we sense that there will not be huge sex differences on this issue. But when we look in the body of the table, we see that two-thirds of the women strongly favor lifting the ban on gay military service, and they are joined by 19 percent more who said simply "favor" (rows 4 and 5 of the table). That's 83 percent in favor! By contrast, the corresponding sum among men is 73 percent, a 10 percentage point difference. Note also that fewer than half of the men strongly favor lifting the ban, whereas more than a quarter simply favor lifting the ban. So there is a difference in the distribution of men and women in the two categories on the favor side. If you look

TABLE 13-14 **Gays in the Military: A Gender Gap?**

Gays Serve in Military?	Male (0)	Female (1)	Totals
(1) Strongly opposed	18.0% (183)	10.8 % (132)	14.1% (315)
(2) Opposed	9.3% (94)	5.8% (71)	7.4% (166)
(3) Favor	28.5% (289)	18.8% (231)	23.2% (520)
(4) Strongly favor	44.2% (449)	64.6% (794)	55.4% (1,244)
Totals	100% (1,015)	100% (1,229)	100% (2,245)

Summary statistics: gamma = .33, tau-*b* = .19, tau-*c* = .21, Somers' *D* = .21, λ = 0, χ^2 = 94.29 with 3 *df*.

at the bottom of the table, a similar conclusion emerges. The ordinal coefficients help a bit. They show first, a modest to weak correlation—as we saw from the percentages—and second, that the relationship is positive.

In this instance, you can think of the variables as having an underlying order. Attitudes toward gays in the military run from low to high support. Gender can be treated as if it were a numeric variable by letting men be 0 and women 1.[9] So as you move across and down the table, going in effect from low values on *X* and *Y* to high values, a slight positive correlation appears. (We place index numbers in parentheses in the table to illustrate the idea, but of course the measures of correlation introduced here do not in any way depend on numerical scale scores.) Beyond saying that there is a limited correlation that the percentages also reveal, these ordinal statistics do not have a common-sense or easily grasped interpretation. The situation improves slightly with the next coefficient.

A Coefficient for Nominal Data

When one or both of the variables in a cross-tabulation are nominal, ordinal coefficients are not appropriate because the identification of concordant and discordant pairs requires that the variables possess an underlying ordering (one value being higher than another). For these tables, different measures of association are employed. Some of the most useful rest on a *proportional-reduction-in-error* interpretation of association. The basic idea is this: You are asked to predict a randomly selected person's category or response level on a variable following two rules. Rule 1 requires you to make the guess in the absence of any other prior information (e.g., predict the individual's position on gun control). The other rule lets you know the person's score on a second variable, which you now take into account in making the prediction (e.g., you now know the individual's gender). Since you are guessing in both situations, you can expect to make some errors, but *if* the two variables are associated, then the using the second rule should lead to fewer errors than following the first.

How many fewer errors depends on how closely the variables are related. If there is no association at all, the expected number of errors should be roughly the same, and the reduction will be minimal. If, on the other hand, the variables are perfectly connected, in the sense that there is a one-to-one connection between the categories of the two variables, you would expect no errors by following rule 2. A "PRE measure" gives the **proportionational reduction in errors**:

$$PRE = \frac{(E_1 - E_2)}{E_1},$$

[9] We could have used any two numbers, such as 1 and 2 or 10 and 21. These numbers don't enter into 'any calculations, but they have marvelous properties in quantitative analysis.

where E_1 is the number of errors made using rule 1 and E_2 is the number made under rule 2.

Suppose for a particular group of subjects the number of rule 1 errors (E_1) predicting variable scores on Y is 500. Now, think about these possibilities:

1. X has no association with Y. Then even using the individuals' X scores, the expected number of errors will still be 500, and the proportional reduction in errors will be (500 – 500)/500 = 0. This is the lower limit of a proportion, and it indicates *no* association.
2. Suppose the categories of X are uniquely associated with those of Y so that if you know X, you can predict Y exactly. The expected number of errors under rule 2 (E_2) will be zero. Consequently, $PRE = (500 - 0)/500 = 1.0$, the upper boundary for the measure. This means *perfect* association (according to this definition).
3. Now, assume that Y and X have a moderate relationship. The expected number of errors following rule 2 might be, say, 200. Now we have

$$PRE = \frac{(500-200)}{500} = \frac{300}{500} = .6.$$

There is then a 60 percent reduction in prediction errors from knowing the value of X, a result that suggests a modest but not complete association.

LAMBDA. Many coefficients of association (e.g., gamma) can be defined in such a way as to lead to a PRE interpretation. We describe only one, however: **Goodman and Kruskal's lambda**. Lambda is a proportional-reduction-in-error coefficient. As we did earlier, imagine predicting a person's score on a variable in the absence of any other information ("rule 1"). What exactly would be the best strategy? If you did not know anything, you might ask what proportion of the population had characteristic A, what proportion characteristic B, and so forth for all of the categories of the dependent variable of interest. Let's say B was the most common (modal) category. Then, without other information, guessing that each individual was a B would produce fewer prediction errors than if you picked any other category. Why? Well, suppose there were 10 As, 60 Bs, and 30 Cs in a population of 100. Select a person at random and guess his or her category. If you picked, say, A, you would on average be wrong 60 + 30 = 90 times out of 100 guesses (90% incorrect). If, on the other hand, you chose C, you would be mistaken 10 + 60 = 70 times (70% errors). Finally, if you guessed the modal (most frequent) category, B, your errors would be on average 10 + 30 = 40. By choosing B (the mode), you do indeed make some incorrect predictions, but many fewer than if you picked any other category. In sum, rule 1 states that, lacking any other data, your best long-run strategy for predicting an individual's class is to choose the modal one, the one with the most observations.

Now suppose you knew each case's score or value on a second variable, *X*. Say you realized a person fell in (or had property) M of the second variable. Rule 2 directs you to look only at the members of M and find its modal category. Assume that category C is most common among those who are misgiven that the observation is an M, guessing C would (over the long haul) lead to the fewest mistakes. So rule 2 simply involves using rule 1 *within* each level of *X*.

The key to understanding lambda, a proportional-reduction-in-error-type measure of association, lies in this fact: *if Y* and *X* are associated, then the probability of making an error of prediction using rule 1 will be greater than the probability of making an error with rule 2. How much greater? The measure of association, lambda (λ), gives the proportional reduction in error:

$$\lambda = \frac{\left(p_{\text{error}1} - p_{\text{error}2}\right)}{p_{\text{error}1}}$$

where $p_{\text{error}1}$ is the probability of making a prediction error with the first rule and similarly $p_{\text{error}2}$ is the likelihood of an error knowing *X*. If the values of *X* are systematically connected to those of *Y*, the errors under the second rule will be less probable than those made under rule 1. In this case, lambda will be greater than zero. In fact, if *no* prediction errors result from rule 2, the probability $p_{\text{error}2}$ will be zero, and

$$\lambda = \frac{\left(p_{\text{error}1} - 0\right)}{p_{\text{error}1}} = \frac{p_{\text{error}1}}{p_{\text{error}1}} = 1.0.$$

But of course if *X* and *Y* are unrelated, then knowing the value of *X* will tell you nothing about *Y*, and in the long run the probability of errors under both rules will be the same. So $p_{\text{error}1} = p_{\text{error}2}$ and

$$\lambda = \frac{\left(p_{\text{error}1} - p_{\text{error}2}\right)}{p_{\text{error}1}} = \frac{(0)}{p_{\text{error}1}} = 0.$$

The upshot is that lambda lies between 0 (no association) and 1.0 ("perfect" association, as defined by the prediction rules). A value of .5 would indicate a 50 percent reduction in errors, which in most situations would be quite a drop and hence suggest a strong relationship. A value of, say, .10—a 10 percent reduction—might signal a weak to nonexistent association. Note that correlation is not an issue here. If there is an *X*-*Y* link of whatever kind, lambda should pick it up. Yet also remember that lambda does *not* take into account the ordering of the categories.

Again, we emphasize the importance of looking at the whole forest (the overall relationship) and not obsessing over a single tree (a measure of association). These kinds of statistics usually depend to a greater or lesser extent on the marginal distributions of the variables. Take care when a preponderance of observations are piled up in one or two categories.[10] For example, the lambda in table 13-13 is .575, which means knowing a person's ideology allows us to predict vote preference reasonably well; we cut prediction errors by more than 50 percent. This result, of course, agrees with our previous conclusion that voting is closely tied to ideology. (If you want to check another of lambda's characteristics, try scrambling the order of the columns in table 13-13. You should get the same result: .575.)

Testing a Cross-Tabulation for Statistical Significance

Before taking up methods for describing relationships between other types of variables, we need to pause to think about this problem. Apart from the hypothetical data, all of the examples presented so far use sample surveys. As samples go, most are quite large with slightly more than 1,000 cases. Nevertheless, since the totals represent only a tiny fraction of the population, one can always ask, "Do observed relationships reflect true patterns, or did they arise from chance or what is called sampling error?" Chapter 12 introduced concepts for answering that sort of question. Here we apply them to cross-classifications.

STATISTICAL INDEPENDENCE. At this point it is useful to introduce a technical term that plays a large role in data analysis and that provides another way to view the strength of a relationship. Suppose we have two nominal or categorical variables, X and Y. For the sake of convenience, we can label the categories of the first a, b, c, . . . and those of the second r, s, t, . . . Let $P(X = a)$ stand for the probability that a randomly selected case has property or value a on variable X, and let $P(Y = r)$ stand for the probability that a randomly selected case has property or value r on Y. These two probabilities are called marginal probabilities and refer simply to the chance that an observation has a particular value (a, for instance) irrespective of its value on another. And, finally, $P(X = a, Y = r)$ stands for the joint probability that a randomly selected observation has *both* property a and property r simultaneously. The two variables are statistically independent if and only if the chances of observing a combination of categories is equal to the marginal probability of one category times the marginal probability of the other:

$$P(X = a, Y = r) = [P(X = a)][P(Y = r)] \text{ for all } a \text{ and } r.$$

[10] Many of these statistics attempt in one way or another to take into account the number of categories and the distribution of cases among them. Going into more detail would take us too far astray.

HOW IT'S DONE

Calculating Lambda

To calculate lambda, follow these steps. (For an example using hypothetical data, see figure 13-4.)

1. Look at the cross-tabulation with both sets of marginal frequency (not percent) totals displayed.
2. Decide which variable is dependent.
3. Find the maximum marginal total for the dependent variable.
4. Subtract this total from table total, *N*, to get errors by method 1: *N* – (maximum frequency) = E_1, the number of predictions errors not knowing the independent variable.
5. In the body of the table, find the maximum frequency within *each* category of the independent variable.
6. Sum the maximums and subtract the total from *N*. Call the result E_2, the number of prediction errors after using knowledge of the independent variable.
7. Calculate lambda:

$$\hat{\lambda} = \frac{(E_1 - E_2)}{E_1}.$$

Note that the numerical value of lambda depends on the choice of independent and dependent variables. Reversing them will usually change $\hat{\lambda}$.

If, for instance, men are as likely to vote as women, then the two variables—gender and voter turnout—are statistically independent because, for example, the probability of observing a male nonvoter in a sample is equal to the probability of observing a male times the probability of picking a nonvoter.

In table 13-15, we see that 100 out of 300 respondents are men and that 210 out of the 300 respondents said they voted. Hence, the marginal probabilities are $P(X = m) = 100/300 = .33$ and $P(Y = v) = 210/300 = .7$. The product of these marginal probabilities is $(.33)(.7) = .23$. Also note that because 70 voters are male, the joint probability of being male *and* voting is $70/300 = .23$, the same as the product of the marginal probabilities. Since the same relation holds for all other combinations in this dataset, we infer that the two variables in table 13-16 are statistically independent.

Now suppose we had the data shown in table 13-16. There the sample consists of 300 respondents, half of whom voted and half of whom did not. The marginal probabilities of voting and not voting are both 150/300 = .5. It is also clear that the marginal probabilities of being upper- and lower-class equal .5. *If* the two variables were statistically independent, the probability that an upper-class respondent voted would be (.5)(.5) = .25. Similarly, the predicted probability (from these marginal totals) that a lower-class individual did not vote would be (.5)(.5) = .25. But we can see from *observed* cell frequencies that actual proportions of upper- and lower-class voters are .33 and .17, respectively. Since the observed joint probabilities do not equal the product of the marginal probabilities, the variables are not statistically independent. Upper-class respondents are more likely to vote than are lower-class individuals.

TABLE 13-15 **Voter Turnout by Gender**

Turnout (*Y*)	Male (m)	Female (f)	Gender (*X*) Total
Voted (*v*)	70	140	210
Did not vote (*nv*)	30	60	90
Total	100	200	300

Note: Hypothetical data. Cell entries are frequencies.

TABLE 13-16 **Voter Turnout by Social Class**

	Social Class (*X*)		
Turnout (*Y*)	Upper (u)	Lower (1)	Total
Voted (*v*)	100	50	150
Did not vote (*nv*)	50	100	150
Total	150	150	300

Note: Hypothetical data. Cell entries are frequencies.

In this context, a test for statistical significance is really a test that two variables in a population are statistically independent. The hypothesis is that in the population, the variables are statistically independent, and we use the observed joint frequencies in a table to decide whether or not this proposition is tenable. Generally speaking, the stronger a relationship is, the more likely it is to be statistically significant, because it is unlikely to arise if the variables are really independent. However, even weak relationships may turn out to be statistically significant in some situations. In the case of cross-tabulations, the determination of statistical significance requires the calculation of a statistic called a chi square, a procedure we discuss next.

CHI-SQUARE TEST FOR INDEPENDENCE. Table 13-17 pertains to civil liberties. It shows by levels of education attainment the degree of agreement with this statement: "Society shouldn't have to put up with those who have political ideas that are extremely different from the majority." The underlying hypothesis is that tolerance of dissent increases with education. By examining the cell proportions and the measures of association, you can surmise that a modest relationship exists between the two variables. (You might reinforce your understanding of the coefficients by interpreting them to yourself.) But is the relationship statistically significant? In the population is there really a relationship between tolerance and education?

TABLE 13-17 **Opinion on Civil Liberties: Tolerance of Dissent**

	Educational Attainment				
Do not put up with extreme differences	**Less than high school**	**High school graduate**	**Some post-high school education**	**College graduate or postgraduate**	**Totals**
Agree	45.58%	41.12%	23.17%	20.45%	31.7% (312)
Uncertain	7.65%	9.24%	6.80%	7.30%	7.8% (77)
Disagree	46.77%	49.63%	70.03%	72.25%	60.5% (596)
Totals	100% (154)	100% (312)	100% (270)	100% (249)	100% (985)

Chi square = 55.66 with 6 *df*.

Gamma = .32, tau-*b* = .20, tau-*c* = .19, Somers' D_{YX} = .24, lambda = 0, φ = 0.24.

Question: "Now I would like to ask about public affairs. Please indicate whether you agree. Society shouldn't have to put up with those who have political ideas that are extremely different from the majority." ("Agree strongly" and "agree" responses have been combined, as have the disagree categories.)

Source: Citizen, Involvement, Democracy Survey, 2006.

Whether or not a relationship is statistically significant usually cannot be determined just by inspecting a cross-tabulation alone; instead, a statistic called **chi square** (χ^2) must be calculated. This statistic essentially compares an observed result—the table produced by sample data—with a "hypothetical" table that would occur if, in the population, the variables were statistically independent. Stated differently, the chi square measures the discrepancy between frequencies actually observed and those we would expect to see if there was no population association between the variables. When each observed cell frequency in a table equals the frequency expected under the **null hypothesis** of independence, chi square will equal zero. Chi square increases as the departures of observed and expected frequencies grow. There is no upper limit to how big the difference can become, but if it passes a certain point—a critical value—there will be reason to reject the hypothesis that the variables are independent.

How is chi square calculated? The observed frequencies are shown in bold in the cross-tabulation in table 13-18. Expected frequencies in each cell of the table are found by multiplying the row and column marginal totals and dividing by the sample size. As an example, consider the first cell in table 13-18. That cell is in the first row, first column of the table, so multiply the row total, 312, by the column total, 154, and then divide by 985, the total sample size in this table. The result is (312 × 154)/985 = 48.78. This is the *expected* frequency in the first cell of the table; it is what we would expect to get in a sample of 985 (with 312 "agrees" and 154 less than high school graduates) *if there is statistical independence in the population.* This is substantially less than the number we actually have, 70, so there is a difference. What about the other cells?

Let's do another example. If there were no association, how many college graduates would we expect to find in the "Disagree" category? Again, find the corresponding marginal totals (here 596 and 249), multiply them, and divide by 985 to get 150.7, the expected number under the null hypothesis. Notice that we keep repeating the phrase "under the . . ." We want to stress that this procedure can be interpreted as measuring the adequacy of a simple model (the model of no association) to these observed data. If the adequacy or fit is good, we say the model partially explains the data, which in turn is a manifestation of the real world. If the assumption of independence is not supported, we wouldn't anticipate that the expected frequencies would equal the observed ones except by chance.

Table 13-17 contains all of the expected frequencies for table 13-18. The overall measure of fit—the observed test statistic—is found by, in effect, comparing observed and expected frequencies. If the sum of differences is relatively small, do not reject the hypothesis of no association. But, if in the aggregate the discrepancy between observed and expected numbers is large, then the model upon which the expected frequencies are calculated is not a summary of the data, and the decision will be to reject the null hypothesis. So what is a large departure from the expected? The statistic is found by subtracting each expected frequency from its observed counterpart, squaring the difference (no minus sign will be left), dividing the quotient by the expected frequency, and then adding the results over all the cells of the table. Hence, for table 13-19 we have

$$\chi^2_{\text{obs}} = \frac{(70-48.8)^2}{48.8} + \frac{(128-98.8)^2}{98.8} + \frac{(63-85.5)^2}{85.5} \ldots + \frac{(180-150.7)^2}{150.7} = 55.66.$$

This observed chi square is 55.66, which we compare to a critical value to help decide whether or not to reject the null hypothesis.

Recall that a statistical hypothesis test entails several steps: specify the null and alternative hypothesis, specify a sample statistic and an appropriate sampling distribution, set the level of significance, find critical values, calculate the observed test

TABLE 13-18 **Observed and Expected Values under Hypothesis of Independence**

Do not put up with extreme differences	Level of Education				Totals
	Less than high school	High school graduate	Some post-high school education	College graduate or postgraduate	
Agree	**70** *48.78*	**128** *98.8*	**63** *85.5*	**51** *78.9*	**312**
Uncertain	**12** *12.0*	**29** *24.4*	**18** *21.1*	**18** *19.5*	**77**
Disagree	**72** *93.2*	**155** *188.8*	**189** *163.4*	**180** *150.7*	**596**
Totals	**154**	**312**	**270**	**249**	**985**

Source: Table 13-17.

Note: Numbers in boldface font are observed frequencies; those in *italics* are expected frequencies under the hypothesis of statistical independence.

statistic, and make a decision. A chi-square test of the **statistical independence** of *Y* and *X* has the same general form.

1. Null hypothesis: *X* and *Y* are statistically independent.
2. Alternative hypothesis: *X* and *Y* are not independent. The nature of the relationship is left unspecified.
3. Sampling distribution: Choose chi square. This distribution is a family, each member of which depends on *degrees of freedom* (*df*). The **degrees of freedom** equals the number of rows (I) minus 1 times the number of columns (J) minus 1 or $(I - 1)(J - 1)$.
4. Level of significance: Choose the probability (α) of incorrectly rejecting a true null hypothesis.
5. Critical value: The chi-square test is always one-tailed. Choose the critical value of chi square from a tabulation to make the critical region (the region of rejection) equal to α.
6. The observed chi square is the sum of the squared differences between observed and expected frequencies, divided by the expected frequency.
7. Reject the null hypothesis if the observed chi square equals or exceeds the critical chi square; that is, reject if $\chi^2_{\text{obs}} \geq \chi^2_{\text{critical}}$. Otherwise, do not reject.

For the tolerance and education example, the null hypothesis is simply that the two variables are independent. The alternative is that they are not. (Yes, this is an uninformative alternative in that it does not specify *how* education and political tolerance might be related. This lack of specificity is a major criticism of the common chi-square test. But this is nevertheless a first step in categorical data analysis.) For this test, we will use $\alpha = .01$ level of significance. To find a critical value, it is necessary to first find the degrees of freedom, which in this case is $(4 - 1)(3 - 1)$, or 6.

Then we look in a chi-square table to find the value that marks the upper 1 percent (the .01 level) of the distribution (see appendix C). Read down the first column (*df*) until you find the degrees of freedom (6 in this case) and then go across to the column for the desired level of significance. With 6 degrees of freedom, the critical value for the .01 level is 16.81. This means that if our observed chi square is greater than or equal to 16.81, we reject the hypothesis of statistical independence. Otherwise, we do not reject it.

The observed chi square for table 13-17 is 55.66 with 6 degrees of freedom. (*Always report the degrees of freedom.*) Clearly, this greatly exceeds the critical value (16.81), so we would reject the independence hypothesis at the .01 level. Indeed, if you look at the chi-square distribution table, you will see that (for 6 degrees of freedom) 55.66 is much larger than the highest listed critical value, 22.46, which defines the .001 level. So really this relationship is "significant" at the .001 level. We place quotation marks around "significant" to reemphasize that all we have done is reject a null hypothesis. We have not necessarily produced a momentous finding. This statement leads to our next point.

The sample size, *N*, and the distribution of cases across the table always have to be taken into account. Large values of chi square occur when the observed and expected tables are quite different and when the sample size upon which the tables are based is large. A weak relationship in a large sample may attain statistical significance, whereas a strong relationship found in a small sample may not. Keep this point in mind. If *N* (the total sample size) is large, the magnitude of the chi-square statistic will usually be large as well, and we will reject the null hypothesis even if the association is quite weak. This point can be seen by looking at tables 13-19 and 13-20. In table 13-19, the chi square of 1.38 suggests that there is virtually no relationship between the categories *X* and *Y*. In table 13-20, which involves a larger sample size but no other difference, the chi-square

TABLE 13-19 Relationship between *X* and *Y* Based on Sample of 300

	Variable *X*			
Variable Y	**A**	**B**	**C**	**TOTAL**
A	30	30	30	90
B	30	30	36	96
C	40	40	34	114
Total	100	100	100	300

$\chi^2 = 1.38$, 4 *df*; $\phi = .07$.

Note: Hypothetical data.

TABLE 13-20 **Relationship between *X* and *Y* Based on Sample of 3,000**

	Variable *X*			
Variable *Y*	**A**	**B**	**C**	**TOTAL**
A	300	300	300	900
B	300	300	360	960
C	400	400	340	1,140
Total	1,000	1,000	1,000	3,000

$\chi^2 = 13.8$, 4 *df*; $\phi = .07$.

Note: Hypothetical data.

statistic (13.8) is now statistically significant (at the .05 level). However, the strength of the relationship between *X* and *Y* is still the same as before—namely, quite small.

The lesson to be drawn here is that when dealing with large samples (say, $N > 1{,}500$), small, inconsequential relationships can be statistically significant.[11] As a result, we must take care to distinguish between statistical and substantive importance. The fact that .chi square rapidly inflates with increases in the sample size has led statisticians to propose measures that try to take *N* into account. A simple one, **phi** (ϕ), adjusts the observed chi-square statistic by dividing it by *N* and taking the square root of the quotient. (Because of the division by *N*, the statistic is sometimes referred to as the "mean square contingency coefficient.") Yet, like chi square, phi does not have a readily interpretable meaning, so it is mostly used for comparison. (In ideal situations, phi varies between 0 and 1, but in many bivariate distributions, it can exceed 1.) We see in tables 13-21 and 13-22 that phi does not change even though the chi-square statistic does. So even though we do not use it much in this book, it comes in handy on occasion. If you look back to table 13-17, you will see that phi = .24, indicating once more the weak to moderate relationship between education and political tolerance.

Generally speaking, the chi-square test is only reliable for relatively large *N*s. Stating exactly how large is difficult because the answer depends on the table's number of rows and columns (or, more formally, its degrees of freedom). Many times, as in

HOW IT'S DONE

The phi Coefficient

Although most software calculates phi as a matter of course, it can be calculated quickly by hand if the observed chi square is available:

$$\pi = \sqrt{\frac{\chi^2_{\text{obs}}}{N}},$$

where *N* is the sample size.

[11] Note, however, that small effects can in some circumstances have theoretical or substantive importance.

a table with many cells, a sample will be large but the table will contain at least some cells with small frequencies. Very few respondents in the CID study reported in Table 13-17 seemed "uncertain," so frequencies in that row are small compared to the others. A rule of thumb directs analysts to be cautious if any cell contains expected frequencies of 5 or fewer, and many cross-classification programs flag these "sparse cells." If you run across this situation, the interpretation of the chi-square value remains the same but should be perhaps advanced with less certainty. Moreover, if the total sample size is less than 20 to 25, alternative procedures are preferable for testing for significance.[12]

Remember: the chi-square statistic in and by itself is not a very good indicator of the strength of an association; rather, it tests the statistical significance of any association that does appear. Assessing relationships is thus a two-step process: (1) measure the strength of the association with percentages, proportions, and coefficients, and (2) test to see if the observed results might have arisen by chance. The first step is the crucial one: make sure the relationship is "worth talking about" and *then* test its significance.

The Relationship between a Categorical Dependent Variable and a Quantitative Variable

Suppose you want to compare the academic performance of students attending charter and public schools. You draw a random sample of student files from a private academy and another sample of student records from a public school of similar size.[13] You thus have an independent variable, "type of school," with two categories, public and charter. Call it X. The dependent variable, Y, is the total score each student receives on a standardized test. In essence, you want to examine the relationship between the categorical variable (X) and a numerical variable (Y). Specifically, you want to know how strong the relationship is and (since it is based on a sample) if it is statistically significant.

The research hypothesis might be that the average (mean) score received by charter school attendees is greater than those of public school attendees. In this case the X-Y relationship can be measured by the *difference of means* of the two groups.

If X has more than two classes, we can compare pairs of means or other more complicated combinations. If we find differences of means of Y among some or all categories of X, then we could argue there is a relationship, its size being determined

[12] See Agresti and Finlay, *Statistical Methods for the Social Sciences*, 264–65, for more information and ideas.about how to proceed.

[13] This research design is obviously too simplistic, but we can use it to demonstrate our point.

by the magnitudes of the differences. If, on the other hand, the means are more or less the same, we might conclude there is no (meaningful) relationship between X and Y.

Difference of Means and Effect Sizes

The difference between one mean and another is an **effect size**, one of the most basic measures of relationships in applied statistics. The name comes from experimental sciences, in which a goal is to measure the effect of a treatment on the dependent variable. A logical measure of an effect is the difference:

$$\Delta = \text{Effect} = \text{Mean of group 1} - \text{Mean of group 2},$$

where Δ (capital Greek letter delta) is the effect size.[14] A logical estimator of Δ is the difference in sample means:

$$\hat{\Delta} = \bar{Y}_{\text{group 1}} - \bar{Y}_{\text{group 2}}.$$

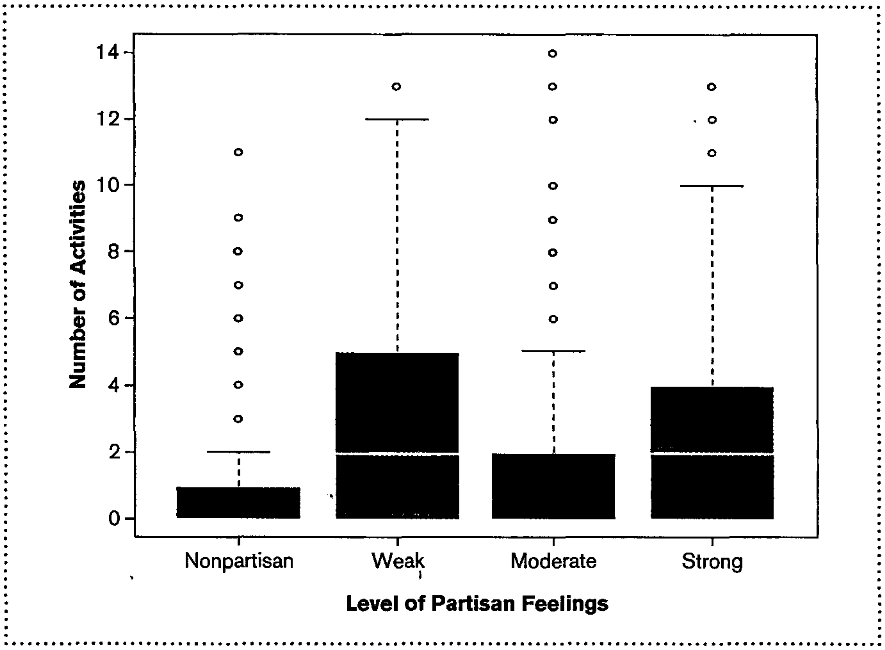
Capital delta with a "hat" is the symbol for the sample estimator of an effect size, and the $\bar{Y}$'s are the sample means for the experimental and control group.

For a change of pace, let's turn to a different substantive question, one that was posed in chapter 1. If justices on the Supreme Court are supposed to follow the law and not their political beliefs, why the hullabaloo over the nomination and confirmation? Shouldn't the best "legal" minds be chosen, regardless of ideology or political affiliation? Of course, everyone knows that Supreme Court decisions are determined by more than just objective interpretation of law and precedent. They surely reflect the political views of the justices as well. After all, presidents nominate justices who share their general philosophy. One way to demonstrate the point is to compare justices' rulings by the party of the nominating president.

Look at figure 13-6, which we will refer to on several occasions. (We have tilted the boxplot on its side to aid in making comparisons.) It displays the voting record of Supreme Court justices nominated and confirmed between 1950 and 2008. (There are twenty-three in all.) Decisions have been limited to those involving union activities,

[14] The quotation marks around "population" are necessary because, technically speaking, there are *no* population experimental and treatment group means. These are hypothetical or theoretical quantities. They could exist only if a researcher could somehow conduct an experiment on an entire population and at the same time treat it as a control. This is a subtle point but one that has far-reaching consequences for how the results of experimental and observational studies are interpreted. An excellent and accessible introduction to this topic is Christopher Winship and Stephen L. Morgan, "The Estimation of Causal Effects from Observational Data," *Annual Review of Sociology* 25 (1999): 659–706. Available at http://www.wjh.harvard.edu/soc/faculty/winship/winship_causal_observational_99.pdf

FIGURE 13-5 **Political Activity by Partisanship**



Source: Citizenship, Involvement, Democracy Survey, 2006.

such as worker safety and labor-management cases. The dependent variable is defined and measured as "the percentage of 'liberal' votes cast by the justice in the area of unions."[15] The point is to identify any meaningful (practical and statistical) differences in behavior between the justices nominated by Republicans and Democrats.

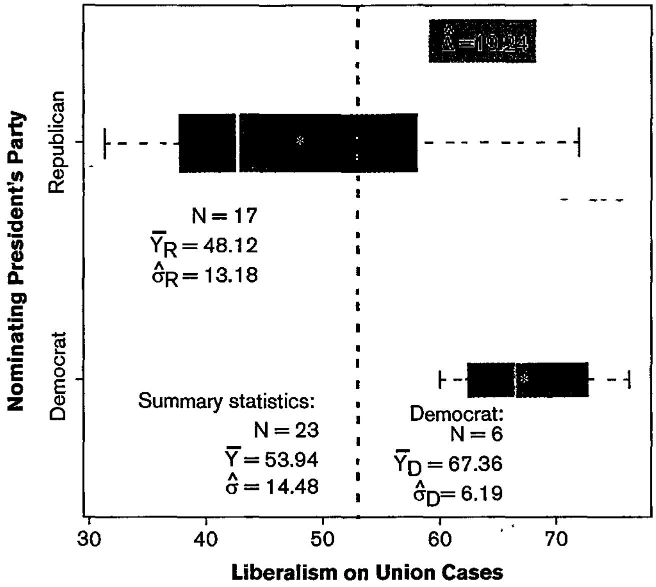
The plot shows a clear difference in the distributions. Besides the medians, which are represented by solid lines in the boxes, we have added the means (67% and 48%). As one might expect, justices selected by Democratic presidents are more liberal on labor issues by approximately 20 percentage points. If you look carefully, you can see that half or more of the "Democrats" score above 65 percent, whereas the same proportion of "Republicans" lie below 45 percent. No Democratic appointee falls below 50 percent liberal on union-related cases. Once more, we see the interconnection between politics and the economy: organized labor does "better" under Democratic than Republican administrations. And we will see later that the stronger the unions are in a country, the less inequality there is. The boxplot shows

[15] Lee Epstein, Thomas G. Walker, Nancy Staudt, Scott A. Hendrickson, and Jason M. Roberts, *Codebook: US Supreme Court Justices Database,* January 26, 2010. Available at http://epstein.usc.edu/research/justicesdata.pdf

more. Note, for instance, that there is more variation among Republican-nominated justices than their Democratic counterparts. Finally, this figure provides an important piece of information that is essential in our further analysis: the number of justices in each group.

A boxplot such as figure 13-6 gives us information useful for conducting tests of significance. The estimation of an effect such as the consequences of party affiliation on judicial decision making requires two *independent* samples of size N_1 and N_2. (If the samples are not independent, alternative statistical procedures have to be used.) The size of the samples also matters because small Ns are handled slightly differently than large ones. As you can see, there are only six Democratic justices and seventeen Republicans.[16] In addition, we have to pay attention to the variation (as measured by the standard deviations) of the two populations from which the samples come. If we can assume that population 1's standard deviation equals population 2's, the test for significance goes in one direction; if we believe the standard

FIGURE 13-6 **Union Liberalism by Nominating President's Party**



Source: Lee Epstein et al., US Supreme Court Justices Database.

[16] Here is an interesting point to consider. The data go back to 1950, yet just eight out of twenty-three justices were nominated by Democratic presidents. It would be fair to say that Republicans have had a chance to influence the nomination process, but the ideological results (e.g., repealing the *Roe v. Wade* decision) have not panned out according to their expectations. It would interesting to think about the reasons why.

deviations are not the same, we follow a different path. Here, we note that the decisions relating to unions of justices nominated by Republican presidents appear to be much less liberal than those of their Democrat-nominated colleagues.

The sample mean union liberalism scale scores are 67.36 percent for the Democrats and 48.12 percent for the Republicans; the difference (the partisan "*effect*") is

$$\hat{\Delta} = \bar{Y}_{\text{Democrat}} - \bar{Y}_{\text{Republican}} = 67.36 - 48.12 = 19.24.$$

The "direction" of the difference, positive, has a nice substantive interpretation: people who are nominated by Democratic presidents tend to vote about 20 percent "more liberal" on union-related cases than do those picked by Republicans.

We have accomplished one objective: estimating the difference between two groups. But is this difference statistically significant? Could it have arisen by chance, especially since the samples are small?[17] Note that once this question is answered, we still have to decide if the observed difference is substantively or politically important. A 20 percent difference might be considered large even though it is based on small samples. Given just these data, we easily see why politicians and interest groups fight tooth and nail over judicial appointments.

The procedures for testing for the significance of a difference of means depend on (among other considerations) sample sizes. We begin with so-called large-sample tests first.

LARGE-SAMPLE DIFFERENCE-OF-MEANS TEST. We are going to compare two sample means, $\bar{Y}_1$ and $\bar{Y}_2$, based on samples of size N_1 and N_2, respectively. In this section, we assume that both *N*s are greater than or equal to 20. The final preliminary point is that the samples have been drawn from populations having means μ_1 and μ_2 and variances σ_1^2 and σ_2^2. For the large-sample tests and confidence intervals, we make no assumptions about these population variances. When it comes time to talk about small samples, we will assume the two variances equal each other.

The judicial dataset is small, so we need a larger one to illustrate the large-sample test. Thus, we shift back to the gender gap hypothesis about male–female differences on current social and political issues. One commonly heard argument is that women are on average a bit more suspicious and hostile toward the military. The 2004 National Election Study survey contains a question that touches on exactly

[17] Needless to say, we do not have samples; the data consist of the entire population. Testing for significance is more of a numerical exercise than an absolute requirement to make generalizations about Supreme Court politics. Nevertheless, for expository purposes we proceed as if we had a simple random sample of justices.

that possibility; it asked respondents to place themselves on a "thermometer" of feelings toward the military, with 0 degrees being coldest or most negative and 100 degrees being most favorable or positive. Presumably, 50 degrees is the neutral position, neither hot nor cold.[18] In the context of American politics, the existence of a gender gap would imply that women have *on average* a slightly less favorable opinion of the military than men do.

Computers perform significance tests practically with the push of a button, but to reinforce your understanding of the logic and assumptions underlying them, we briefly outline the test procedure. The null hypothesis is that thermometer scores are *equal,* or symbolically (letting μ's represent population means),

$$H_0: \mu_{\text{male}} = \mu_{\text{female}} \textit{ or } H_0: \mu_{\text{male}} - \mu_{\text{female}} = 0 \textit{ or } H_0: \Delta_{\mu\text{male}-\mu\text{female}} = 0.$$

Here we have stated the null hypothesis in three equivalent ways. The alternative hypothesis is that women have on average lower thermometer scores, and it too can be written several ways:

$$H_A: \mu_{\text{female}} < \mu_{\text{male}} \textit{ or } H_A: \mu_{\text{female}} - \mu_{\text{male}} < 0 \textit{ or } H_0: \Delta_{\mu\text{female}-\mu\text{male}} < 0.$$

In essence, we are testing whether the population difference of means is zero or negative. Since only values much less than zero are of interest, this is a one-sided test. Let us test at the .01 level of significance, which means that if we should reject the null hypothesis, we may be making a type I error (falsely rejecting H_0), but the chances of doing so are 1 in 100 (.01).

A sample difference of means based on large samples has a normal distribution, so to find an appropriate critical value for testing the null hypothesis, we use the tabulated standard normal distribution or z distribution in appendix A. The critical value has to be chosen in such a way as to make the probability of a type 1 error equal to .01. Recall that the table gives z scores that cut off the upper α proportion of the distribution. The critical value that cuts off .1% (.01) of the area under the normal curve is 2.325. (We interpolated between 2.32 and 2.33.) Any observed test statistic greater than or equal to this value will lead to the rejection of H_0; if it is less, the null hypothesis still stands.

[18] The actual question is, "I'll read the name of a person [or institution] and I'd like you to rate that person using something we call the feeling thermometer. Ratings between 50 degrees and 100 degrees mean that you feel favorable and warm toward the person. Ratings between 0 degrees and 50 degrees mean that you don't feel favorable toward the person and that you don't care too much for that person. You would rate the person at the 50 degree mark if you don't feel particularly warm or cold toward the person. . . . Still using the thermometer, how would you rate the following. . . ." (American National Election Study [ANES], *2004 HTML Codebook,* July 14, 2006; The American National Election Studies, http://www.electionstudies.org/; The 2004 National Election Study [dataset]. Ann Arbor: University of Michigan, Center for Political Studies [producer and distributor]).

For this test, the effect, $\hat{\Delta}$, is converted to an observed test statistic, z, by the general formula:

$$z_{\text{obs}} = \frac{\text{Estimated difference } (\hat{\Delta})}{\text{Estimated standard error}}.$$

The standard error in the denominator is the variation of the sampling distribution. The z_{obs} is compared to the critical z to reach a decision: If $|z_{\text{obs}}| \geq z_{\text{critical}}$, reject H; otherwise, do not reject.

The estimated effect size, $\bar{Y}_{\text{male}} - \bar{Y}_{\text{female}} = 79.56 - 80.07 = -.52$, is not much to write home about, but is it statistically significant? The test results appear in table 13-21.

The observed test statistic ($z_{\text{obs}} = -.393$) is considerably less than the critical value. In view of these data and the way we framed the problem, there is no reason to reject the null hypothesis. More important, the conclusion is that on this issue measured in this way, there appears to be no gender gap. This finding is not, of course, the end of the story. Some of the response variables studied in the section on cross-tabulations *did* reveal a modest male–female divergence on a couple of issues. But it does suggest that gender politics in America may be more complex than conventional wisdom might indicate. (Also bear in mind the timing of the survey, October to December 2004. The war in Iraq was just a year old, and the shock of 9/11 had not dissipated.)

SMALL-SAMPLE TESTS OF THE DIFFERENCE OF MEANS. The preceding test assumed relatively large sample sizes. What about smaller ones? We can return to the Supreme Court data in which the subsample sizes are $N_D = 6$ and $N_R = 17$—both less than 30, our arbitrary cutoff point for deciding what is small. Given the sample sizes and unequal variances, we have to adjust the test procedures. As in the case of single samples (see chapter 11), when dealing with small *N*s, we apply the *t* distribution instead of the standard normal (z) distribution. Under certain assumptions, the difference of means approximately follows a *t* distribution with degrees of freedom, *df*, that are a function of the sample sizes.

All of these considerations lead to two methods for testing whether one μ differs from another:

- Method I (Student's *t*-test): Assume variances are equal $(\sigma_1^2 = \sigma_2^2)$
  - $df = N_1 + N_2 - 2$.
- Method II (Welch *t*-test): Assume unequal variances $(\sigma_1^2 \neq \sigma_2^2)$.
  - *df* is more complicated and we leave its calculations to computer software, since most of the testing you will be doing will be with the help of statistical software.

TABLE 13-21 **Large-Sample Difference-of-Means Test (Gender by Attitudes toward the Military)**

Gender	Sample Size	Mean	Standard Deviation	Standard Error of Mean
Male	518	79.56	20.59	.905
Female	531	80.07	21.99	.954

$\hat{D} = -.52$.

$z_{\text{obs}} = -.393$, p-value = .614.

Confidence interval for difference of mean: −3.913, 2.878.

Source: 2004 National Election Study.

Both cases assume independent random sampling;[19] the dependent variable may or may not be normally distributed.[20]

A quick measure of differences in variation is the "variance ratio":

$$\frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2}.$$

If the ratio is about 1, the variances are roughly equal; if they differ, the ratio will be less than zero or greater than 1.[21] For the current data, the variance of Republican-nominated justices is about four times that of their Democratic counterparts (173.60 versus 38.34). We'll ignore the discrepancy for the moment by assuming equal variances and then later apply method II to determine whether our conclusion changes when we are more rigorous.

The null hypothesis has the usual form:

$$H_0 : \mu_{\text{Democrat}} = \mu_{\text{Republican}} \; or \; H_0 : \mu_{\text{Democrat}} - \mu_{\text{Republican}} = 0 \; or \; H_0 : \Delta_{\mu_{\text{Democrat}} - \mu_{\text{Republican}}} = 0.$$

[19] There is a ton of literature on the analysis of "paired" or matched samples in which members of one group are selected to match some characteristics of members of the other one. See Agresti and Finlay, *Statistical Methods for the Social Sciences,* 226–32.

[20] Nonnormality seems to be most troublesome when one or both subpopulation distributions are skewed.

[21] There are tests for equality of variances, but none performs especially well in all circumstances. See Agresti and Finlay, *Statistical Methods for the Social Sciences,* 220–24. For a more technical discussion, see Richard J. Larsen and Morris L. Marx, *An Introduction to Mathematical Statistics and Its Applications* (Englewood Cliffs, N.J.: Prentice-Hall, 1981), 329–33.

The alternative is that decision making differs by party and so the means will not equal:

$$H_A\!:\mu_{\text{Democrat}} \neq \mu_{\text{Republican}}$$

The choice of the alternative hypothesis dictates a two-tailed test (that is, large departures from the hypothesized difference in *either* direction will be grounds for rejecting H_0).

We need three formulas for this test: (1) the "pooled" standard deviation of both samples, (2) the standard error, and finally, (3) the observed:

1. Pooled standard deviation: $\hat{\sigma}_P = \sqrt{\dfrac{(N_1 - 1)\hat{\sigma}_1^2 + (N_2 - 1)\hat{\sigma}_2^2}{(N_1 + N_2 - 2)}}.$

2. Standard error: $\hat{\sigma}_{\bar{Y}_1 - \bar{Y}_2} = \hat{\sigma}_P \sqrt{1/N_1 + 1/N_2}.$

3. Observed test statistic: $t_{\text{obs}} = \dfrac{(\bar{Y}_1 - \bar{Y}_2)}{\hat{\sigma}_{\bar{Y}_1 - \bar{Y}_2}}.$

Using these formulas, the calculations for the Supreme Court data are

$$\hat{\sigma}_p = \sqrt{\frac{(6-1)(6.19)^2 + (17-1)(13.18)^2}{6+17-2}} = \sqrt{\frac{2970.88}{21}} = 11.89.$$

Therefore, the standard error turns out to be

$$\hat{\sigma}_{\bar{Y}_1 - \bar{Y}_2} = 11.89\sqrt{\frac{1}{6} + \frac{1}{17}} = 5.65,$$

and the observed *t* is

$$t_{obs} = \frac{19.24}{5.65} = 3.41.$$

The correct degrees of freedom for method I is $df = N_1 + N_2 - 2$, or in this case $6 + 17 - 2 = 21$. For $\alpha = .01$, the critical *t* is 2.51. Since $t_{\text{obs}} > t_{\text{crit}}$, we reject H_0.

So the decision is to reject the null hypothesis that Supreme Court justices nominated by Democratic presidents have the same average union cases scores as those supported by Republicans.

HOW IT'S DONE

Calculate Small-Sample Difference-of-Means Tests

Required sample statistics are as follows:

- N_1 and N_2, the sample sizes
- σ_1^2 and σ_2^2, the sample variances
- $\bar{Y}_1$ and $\bar{Y}_2$, the sample means

Assume $N_1 + N_2 \leq 30$. (Pay attention to subscripts.)

Method I: Variances Equal $(\hat{\sigma}_1 = \hat{\sigma}_2)$

Step 1: Pooled estimator of common variance:

$$\hat{\sigma}_P = \sqrt{\frac{(N_1 - 1)\hat{\sigma}_1^2 + (N_2 - 1)\hat{\sigma}_2^2}{(N_1 + N_2 - 2)}}.$$

Step 2: Estimated standard error of difference

$$\hat{\sigma}_{\bar{Y}_1 - \bar{Y}_2} = \hat{\sigma}_P \sqrt{{}^{1}\!/_{N_1} + {}^{1}\!/_{N_2}}.$$

Step 3: Observed t

$$t_{\text{obs}} = \frac{(\bar{Y}_1 - \bar{Y}_2)}{\hat{\sigma}_{\bar{Y}_1 - \bar{Y}_2}}.$$

Step 4: Degrees of freedom

$$df = N_1 + N_2 - 2.$$

Method II: Variances Unequal $(\sigma_1 \neq \sigma_2)(N_1$ and N_2 not necessarily equal)

Step 1: Estimated standard error of difference

$$\hat{\sigma}_{\bar{Y}_1 - \bar{Y}_2} = \sqrt{\frac{\hat{\sigma}_1^2}{N_1} + \frac{\hat{\sigma}_2^2}{N_2}}.$$

Step 2: Observed t

$$t_{\text{obs}} = \frac{(\bar{Y}_1 - \bar{Y}_2)}{\hat{\sigma}_{\bar{Y}_1 - \bar{Y}_2}}.$$

Step 3: Degrees of freedom

$$df = \frac{\left(\frac{\hat{\sigma}_1^2}{N_1} + \frac{\hat{\sigma}_2^2}{N_2}\right)^2}{\left(\left(\frac{\hat{\sigma}_1^2}{N_1}\right)^2 \Big/ (N_1 - 1)\right) + \left(\left(\frac{\hat{\sigma}_2^2}{N_2}\right)^2 \Big/ (N_2 - 1)\right)}.$$

How much have we learned from this exercise? We observed a difference in judicial behavior, 19.23 percent, which we judge to be statistically significant. This analysis supplies a statistical basis for the argument that the Supreme Court is a politicized institution, just like Congress and the president.

METHOD II: UNEQUAL SUBPOPULATION VARIANCES (WELCH *T*-TEST). If the subpopulation variances are not equal, the *t*-test performs rather "poorly," in that levels of significance may be erroneous. (Having equal sample sizes, N_1 and N_2, helps, but we use a modification of method I to make the test when variances are dissimilar.) The procedure generally follows the other means tests, but we have to adjust the degrees of freedom and the standard error. To find the degrees of freedom for the situation we're in—the subpopulation variances differ—the calculations are not as straightforward as before. Fortunately, most software computes the degrees of freedom, so we do not dwell on them here. Instead we concentrate on what the test results tell us about judicial behavior.[22] It is where the terms have been defined earlier. When the computer finishes crunching the numbers, the *df* for the small-sample test in this case is 18.77.

Yes, it's a strange degrees of freedom but perfectly proper. In the absence of a computer, we can treat this as approximately 18 and use appendix B to find the corresponding critical value, which is 2.55 (for a one-tailed test at the .01 level). Any observed t value greater than or equal to this number will lead to the rejection of the null hypothesis in favor of the alternative.

The observed test statistic closely resembles the one calculated for large samples. The general form is the hypothesized effect, Δ, subtracted from the observed effect, $\hat{\Delta}$, or the difference in the corresponding subpopulation means, divided by the estimated standard error of $\hat{\Delta}$. The standard error is just the weighted average of the subpopulation variances:

$$\hat{\sigma}_{\bar{Y}_1-\bar{Y}_2} = \sqrt{\frac{\hat{\sigma}_1^2}{N_1} + \frac{\hat{\sigma}_2^2}{N_2}}.$$

The test statistic is for these data is

$$t_{obs} = \frac{(\bar{Y}_1 - \bar{Y}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\hat{\sigma}_1^2}{N_1} + \frac{\hat{\sigma}_2^2}{N_2}}} = \frac{\hat{\Delta}}{\hat{\sigma}_{\bar{Y}_1-\bar{Y}_2}}.$$

[22] The formula *df* is

$$df = \frac{\left(\frac{\hat{\sigma}_1^2}{N_1} + \frac{\hat{\sigma}_2^2}{N_2}\right)^2}{\left(\left(\frac{\hat{\sigma}_1^2}{N_1}\right)^2 \Big/ (N_1 - 1)\right) + \left(\left(\frac{\hat{\sigma}_2^2}{N_2}\right)^2 \Big/ (N_2 - 1)\right)}.$$

Here is the calculation of the observed difference, its estimated standard error, and the observed *t* all in one fell swoop (you can get the values from the figure):

$$t_{obs} = \frac{(67.36 - 48.13) - (0)}{\sqrt{\dfrac{6.19^2}{6} + \dfrac{13.18^2}{17}}} = \frac{19.23}{4.08} = 4.71.$$

The formula for the standard error looks complicated, but it simply tells us to square the standard deviations in each subpopulation, divide by the respective *N*s, add the two quotients, and take the square root. Incidentally, the simplification of the numerator is possible because $\mu_{\text{Democrat}} - \mu_{\text{Republican}}$ is hypothesized to be zero and $\hat{\Delta} = \hat{\Delta} = \bar{Y}_{\text{Democrat}} - \bar{Y}_{\text{Republican}}$. Table 13-22 presents the test results.

The observed *t* (4.71) easily exceeds the critical value (2.55)—in fact, it is greater than all *t*s with 18 *df*—so we reject the null hypothesis at the .05 level and indeed at the .001 level. As a reminder, the phrase "*p*-value = .000" at the bottom of table 13-22 is the attained probability of the observed *t*. This means that *if* the null hypothesis of no difference of means is true, we have found a very unusual result, one with a probability is less than .001 or 1 in 1,000. The observed *t* is close to the value achieved according to method I (4.71 versus 3.38), so the political conclusion is the same as the one derived from method I.

CONFIDENCE INTERVALS. Chapter 12 underscored the value of confidence intervals. Remember, confidence intervals are lower and upper boundaries that probably enclose the population value of an estimator. Go back to table 13-21, which reports the 99 percent confidence intervals for the estimated gender effect. The intervals extend from –3.913 to 2.878. Thus, 99 times out of 100, the intervals will include the population difference of means. In

TABLE 13-22 Small-Sample Difference-of-Means Test

Nominating Party	Sample Sizes	Mean	Standard Deviation
Democrat	6	67.36	6.19
Republican	17	48.13	12.78

$\hat{\Delta} = 19.23$.

Method I: $\hat{\sigma}_{\bar{Y}_1-\bar{Y}_2} = 5.65$; $t_{\text{obs}} = 3.41$; $df = 21$; $t_{\text{crit}} = 2.51$; *p*-value = .000.

99% confidence intervals: 3.80 – 34.67.

Method II: $\hat{\sigma}_{\bar{Y}_1-\bar{Y}_2} = 4.08$; $t_{\text{obs}} = 4.71$; $df = 18.77$; $t_{\text{crit}} = 2.55$; *p*-value = .000.

other words, we are 99 percent certain that the difference in men and women's mean thermometer scores lies between –3.913 and 2.878. Since the intervals extend from about –4 to 3, we have reason to believe that zero is a possible value of the population difference of means. What is more, because these limits are based on the same α level used in the test of significance, we have in effect tested the null hypothesis a second way: by observing that the 99 percent confidence intervals included the (null) hypothesized value of zero, we do not reject H_0. The substantive interpretation is thus that there is no difference between men and women on this issue.

CALCULATING CONFIDENCE INTERVALS. Although you will undoubtedly use packaged statistical software to find confidence intervals, the computation is not difficult. The general form is

Estimated difference of means ± critical value × estimated standard error.

In words, find the critical value appropriate for the alpha level, multiply it by the standard error, and then add and subtract the product to the difference to get the upper and lower bounds. The critical value is the *z* or *t* used in a hypothesis test (at the desired α level), and the standard error is the standard error of the difference of means. The precise numbers will depend on sample sizes and on whether or not equal population variances are assumed. For those who want a more precise formulation, confidence intervals for a difference of means or effect are

$$\hat{\Delta} \pm \delta_\alpha\left(\hat{\sigma}_{\bar{Y}_1-\bar{Y}_2}\right),$$

where $\hat{\Delta}$ is the estimated effect, the δ_α is a critical value for $(1 - \alpha)$ percent confidence intervals, and $\hat{\sigma}_{\bar{Y}_1-\bar{Y}_2}$ is the estimated standard error of the difference.

There are two situations.

LARGE-SAMPLE INTERVALS. If the *N*s are greater than 20, the confidence intervals are

$$\hat{\Delta} + z_{\alpha/2}\hat{\sigma}_{\bar{Y}_1-\bar{Y}_2} \textit{ and } \hat{\Delta} - z_{\alpha/2}\hat{\sigma}_{\bar{Y}_1-\bar{Y}_2}.$$

Example: Assume $N_1 = N_2 = 100$ and sample standard deviations $\hat{\sigma}_1 = 5$ and $\hat{\sigma}_2 = 4$. If $\bar{Y}_1 = 50$ and $\bar{Y}_2 = 40$, the estimated difference of means is $\hat{\Delta} = 50 - 40 = 10$. The formula for estimated standard error of $\hat{\Delta}$ is

$$\hat{\sigma}_{\bar{Y}_1-\bar{Y}_2} = \sqrt{\frac{\hat{\sigma}_1^2}{N_1} + \frac{\hat{\sigma}_2^2}{N_2}}.$$

HOW IT'S DONE

Large-Sample Confidence Intervals for a Difference of Means

If N_1 and $N_2 \geq 20$, the $(1-\alpha)$ percent confidence intervals for $\hat{\Delta} = \bar{Y}_1 - \bar{Y}_2$ are

$$\hat{\Delta} \pm z_{\text{critical}} \hat{\sigma}_{\bar{Y}_1 - \bar{Y}_2},$$

where

$$\hat{\sigma}_{\bar{Y}_1 - \bar{Y}_2} = \sqrt{\frac{\hat{\sigma}_1^2}{N_1} + \frac{\hat{\sigma}_2^2}{N_2}}$$

For 95 percent confidence intervals, the appropriate critical z is 1.96. (Why?[23]) Hence, the upper limit is

$$\begin{aligned} 10 + 1.96\sqrt{\frac{5^2}{100} + \frac{4^2}{100}} &= 10 + 1.96\sqrt{.25 + .16} \\ &= 10 + 1.96\sqrt{(.41)} \\ &= 10 + 1.96\,(.64) \\ &= 10 + 1.255 \\ &= 11.255, \end{aligned}$$

and the lower limit is

$$\begin{aligned} 10 - 1.96\sqrt{\frac{5^2}{100} + \frac{4^2}{100}} &= 10 - 1.96\sqrt{.25 + .16} \\ &= 10 - 1.96(.64) \\ &= 10 - 1.255 \\ &= 8.745. \end{aligned}$$

The procedure we used to construct the confidence intervals has a 95 percent chance of including the population difference of means. Loosely speaking, we are 95 percent certain that the interval 8.745 to 11.255 contains the true value of $\mu_1 - \mu_2$.

[23] Because that value cuts off the upper .05/2 = .025 portion of the standard normal distribution. Check appendix A.

SMALL-SAMPLE INTERVALS. Intervals for smaller samples (the *N*s are less than 20 or so) have the same general form, except that a critical *t* with approximately $N_1 + N_2 - 2$ (or as calculated previously with the formula) degrees of freedom replaces the critical *z* value. And we use the appropriate standard error depending on the assumption about equal variances.

$$\hat{\Delta} + t_{\alpha/2,N_1+N_2-2}\,\hat{\sigma}_{\overline{Y}_1-\overline{Y}_2} \textit{ and } \hat{\Delta} - t_{\alpha/2,N_1+N_2-2}\,\hat{\sigma}_{\overline{Y}_1-\overline{Y}_2}.$$

Example: As before, let $\overline{Y}_1 = 50$ and $\overline{Y}_2 = 40$ (hence, $\hat{\Delta} = 10$) and $\hat{\sigma}_1 = 5$ and $\hat{\sigma}_2 = 4$. This time, however, set $N_1 = N_2 = 10$. Assume first equal population variances. The pooled estimated of the population standard deviation is

$$\hat{\sigma}_{\text{pooled}} = \sqrt{\frac{(10-1)5^2 + (10-1)4^2}{10+10-2}} = \sqrt{\frac{(9)(25)+(9)(16)}{18}} = \sqrt{\frac{369}{18}} = 4.53,$$

from which we find the estimated standard error to be

$$\hat{\sigma}_{\overline{Y}_1-\overline{Y}_2} = 4.53\sqrt{\frac{1}{10}+\frac{1}{10}} = 2.02.$$

We want 95 percent confidence intervals, so the necessary critical *t* is 2.101 (look in appendix B, the row for 18 *df* and $t_{.025}$). The upper-limit interval turns out to be

$$10 + (2.10)(2.02) = 10 + 4.25 = 14.25,$$

whereas the lower limit is

$$10 - (2.10)(2.02) = 10 - 4.25 = 5.75.$$

As an aside, here are the intervals using the second method. First, the standard error:

$$\begin{aligned}\hat{\sigma}_{\overline{Y}_1-\overline{Y}_2} &= \sqrt{\frac{\hat{\sigma}_1^2}{N_1}+\frac{\hat{\sigma}_2^2}{N_2}}\\ &= \sqrt{\frac{5^2}{10}+\frac{4^2}{10}}\\ &= \sqrt{2.5+1.6}\\ &= 2.02.\end{aligned}$$

The degrees of freedom works out to be 16.81. The appropriate critical t for the .01 level of significance, which corresponds to 99 percent confidence intervals, is 2.898. The intervals for the estimated difference ($\hat{\Delta}=10$) are

$$10-(2.898)(2.02)=4.13 \text{ and } 10+(2.898)(2.02)=15.87.$$

These intervals are considerably wider than the previous ones because of the much smaller samples. (Remember from chapter 12 we stressed that, other things being equal, the larger the samples, the smaller the confidence intervals.)

HOW IT'S DONE

Small-Sample Confidence Intervals for a Difference of Means

Let N_1 and N_2 be the sizes of samples 1 and 2, respectively. Assume if N_1 and $N_2 < 20$.

$(1-\alpha)$ percent confidence intervals for $\hat{\Delta}=\bar{Y}_1-\bar{Y}_2$ are

$$\hat{\Delta} \pm t_{\alpha/2,N_1+N_2-2}\hat{\sigma}_{\bar{Y}_1-\bar{Y}_2}.$$

Method I: σ_1 and σ_2 are equal.

t_{crit} is the t value with N_1+N_2-2 degrees of freedom for the chosen α level of significance and $\hat{\sigma}_{\bar{Y}_1-\bar{Y}_2}$, the estimated standard error of the sample difference of means:

$$\hat{\sigma}_{\bar{Y}_1-\bar{Y}_2}=\hat{\sigma}_{\text{pooled}}\sqrt{\frac{1}{N_1}+\frac{1}{N_2}},$$

and $\hat{\sigma}_{\text{pooled}}$ is the pooled estimator of the common population standard deviation:

$$\hat{\sigma}_{\text{pooled}}=\sqrt{\frac{(N_1-1)\hat{\sigma}_1^2+(N_2-1)\hat{\sigma}_2^2}{N_1+N_2-2}}.$$

Method II: σ_1 and σ_2 are not equal.

t_{crit} is the t value with degrees of freedom calculated as

$$df=\frac{\left(\frac{\hat{\sigma}_1^2}{N_1}+\frac{\hat{\sigma}_2^2}{N_2}\right)^2}{\left(\left(\frac{\hat{\sigma}_1^2}{N_1}\right)^2\Big/(N_1-1)\right)+\left(\left(\frac{\hat{\sigma}_2^2}{N_2}\right)^2\Big/(N_2-1)\right)}.$$

The appropriate standard error is

$$\hat{\sigma}_{\bar{Y}_1-\bar{Y}_2}=\sqrt{\frac{\hat{\sigma}_1^2}{N_1}+\frac{\hat{\sigma}_2^2}{N_2}}.$$

To wrap up, refer first to table 13-21 and note the confidence intervals for the difference of the means between men's and women's scale scores on feelings toward the military. They stretch from –3.913 to 2.878. That is, the difference in means might be negative, meaning women have less favorable attitudes than men, but it could also be positive, meaning that women would feel more favorable toward the military than men. Yet, another possible value is zero. A zero difference, of course, disconfirms the gender gap hypothesis. This conclusion matches the one we made on the basis of just the difference-of-means test itself. This equivalence results from the close connection between hypothesis testing and interval estimation.

Difference of Proportions

Closely related to an analysis of differences of means is the comparison of proportions. Testing for a difference of proportions follows exactly the same steps as the previous tests except for relatively minor adjustments in formulas. When the goal is to measure the difference between two sample proportions, p_1 and p_2, and the data come from two independent samples, place confidence intervals around the estimated difference and test the hypothesis that the population difference of proportions, $P_1 - P_2$, equals a specific value, usually zero. This test has all of the elements of difference-of-means tests, which should not be surprising because we can interpret proportions as a kind of mean. Thus, to check the significance of a difference of proportions, we need the hypotheses (e.g., H_0: $\Delta_p = P_1 - P_2 = 0$ and H_A: $\Delta_p = P_1 - P_2 \neq 0$), decision rules (e.g., α–level = .01, two-sided test, and corresponding critical value), the estimated difference (i.e., $\hat{\Delta}_{p_1 - p_2}$), degrees of freedom (if samples are small), and the standard error. For a difference of proportions test, the standard error is

$$\hat{\sigma}_{p_1 - p_2} = \sqrt{p * (1 - p*)\left[\frac{1}{N_1} + \frac{1}{N_2}\right]},$$

where p^* is the overall sample proportion in the comparison and the Ns are the respective subsample sizes. For confidence intervals, use

$$\hat{\sigma}_{p_1 - p_2} = \sqrt{\frac{(p_1)(1 - p_1)}{N_1} + \frac{(p_2)(1 - p_2)}{N_2}}.$$

The test statistic has the same general form as the one for the difference of means. For large samples,

$$z_{\text{obs}} = \frac{(p_1 - p_2) - (P_1 - P_2)}{\hat{\sigma}_{p_1 - p_2}}.$$

To motivate the discussion, let us stick with the gender gap problem. The 2004 NES survey asked respondents to rate various groups on "feeling thermometers." This time we take attitudes toward "feminists" as the dependent variable, but instead of comparing means, we will investigate the differences in the proportions of men and women who rate feminists "negatively." Thermometer ratings of less than 50 are considered "unfavorable" ratings, while scores of 50 and higher are considered "favorable." (This is admittedly an arbitrary dividing line, and we use it merely to illustrate the method. As a rule of thumb, one should apply the statistical technique that preserves the level of measurement.)

Table 13-23 summarizes the data and analysis results. For large samples, a difference of proportion has a normal distribution with mean $P_1 - P_2$ and standard error $\sigma_{p_1 - p_2}$. As a consequence, we compute an observed z to compare with the critical value. For the sake of argument, we choose a two-tailed test at the .05 level. The null hypothesis is therefore H_0: $P_1 = P_2$ or H_0: $P_1 - P_2 = 0$, and the alternative is simply H_A: $P_1 \neq P_2$. The sample sizes ($N_{\text{Male}} = 496$, $N_{\text{Female}} = 518$) are quite large, so the z distribution is appropriate, and the critical value for the test is 1.96.

The estimated proportions are the cell frequencies divided by the totals (e.g., $p_{\text{Male}} = 139/496 = .280$).

Usually (as in the present case) the null hypothesis is simply H_0: $P_1 - P_2 = 0$, and the last term in the numerator drops out. Notice that because the observed test statistic (3.12) is considerably larger than 1.96, the hypothesis that the population proportions are the same is rejected at the $\alpha = .05$ level. In addition, we report

TABLE 13-23 Difference of Proportions Test (Gender by Attitudes toward "Feminists")

	Gender		
Estimated Proportion	**Male**	**Female**	**Total**
Proportion negative	.280 (139)	.197 (102)	.238 (241)
Proportion positive	.720 (357)	.803 (416)	.762 (773)
Total	1.0 (496)	1.0 (518)	1.0 (1,014)

Estimated difference, $P_1 - P_2$: .280 – .197 = .083.

$z_{\text{obs}} = 3.12$; $z_{\text{crit}} = 1.96$.

Since $z_{\text{obs}} > 1.96$, reject H_0 at .05 level; p-value < .0014.

Confidence interval: .031, .136.

Source: 2004 National Election Study.

that the attained probability of this z is less than .0014, further confirming the decision. In two words, the difference in sample proportions is "statistically significant." There does appear to be a gap in attitudes toward feminists between men and women.

Before overinterpreting the result, however, note the 95 percent confidence intervals (CI) in table 13-23. They extend from about .03 to .14. In other words, the male–female split could be as small as .03, a difference that might have little if any practical importance. Moreover, the upper limit, .14, may not be large for practical purposes.

We conclude with a methodological and a substantive lesson. We have evidence of a small gender gap in these and previously reported data. Yet it does not seem earth shattering, and is not evidence of a major divide in American politics. When it comes to attitudes and partisanship, the gender gap probably pales into insignificance compared with other divisions in society such as racial, regional, class, or ethnic cleavages. The point about statistics is this: examine data from several perspectives. Besides considering the significance of a result measure, think about its magnitude. Once again we see the value of confidence intervals. In this case, they tell us that there might be a modest gap in feelings, but a trivial one is a possibility as well.

In this procedure, the dependent variable is quantitative, and an important measure of the variation is the sum of squared deviations from the mean. To make this concept understandable, consider a variable whose mean is 10. Again, *effect size* means pretty much what the name says: it is a numerical indicator of the impact of an independent variable on a dependent variable. We do not want to get too far ahead, but empirical propositions and theories often stand or fall on effect sizes, not on whether they are "statistically significant."

Analysis of Variance (ANOVA)

Analysis of variance, or ANOVA, extends the previous method to the comparison of more than two means. As before, the dependent variable (Y) is quantitative. The independent or explanatory variable (X), sometimes referred to as a treatment or factor, consists of several categories. This procedure treats the observations in the separate categories as independent samples from populations. If the data constitute random samples (and certain other conditions are met), you apply ANOVA to test a null hypothesis such as H_0: $\mu_1 = \mu_2 = \mu_3$ . . . and so on, where the μs are the population means of the groups formed by X.

Suppose, for example, that you have a variable (X) with three categories—A, B, and C—and a sample of observations within each of those categories. For each observation, there is a measurement on a quantitative dependent variable (*Y*). Thus, within every category or group, you can find the mean of *Y*. ANOVA digs into such data to discover (1) if there are any differences among the means, (2) which specific means differ and by how much, and (3) assuming the observations are sampled from a population, whether the observed differences have arisen by chance or reflect real variation among the categories or groups in *X*.

EXPLAINED AND UNEXPLAINED VARIATION. The inclusion of the word *variance* in "analysis of variance" may throw you. If the procedure deals with means, why not call it the "analysis of means"? As we said in chapter 2 and elsewhere, the goal of empirical research is to explain differences. In statistics, the variation from all sources is frequently called the total variation. In a sample or observed batch of data, the total variation of a variable is measured by the total sum of squares, which is the summation of the squared deviation of each observation from its mean. Symbolically,

$$TSS = \sum_{i=1}^{N} \left(Y_i - \bar{Y}\right)^2.$$

Identified and measured variables explain some of this overall variation; the explained part is called, naturally, the explained variation. What's left over is the unexplained variation. Figure 13-7 may help clarify this point. It shows the data in table 13-24 as a dot chart of individual values. Notice first the considerable variation among the points. By looking at the graph carefully, you may see two kinds of variation. For example, the members of group or category A differ from members of categories B and C. But these observations also vary among themselves. In all three groups, four out of five observations lie above or below their category means. (The mean in A, for instance, is 14, and two scores are above and below it.)

In ANOVA parlance, two types of variation add up to the overall variation, or **total variance**. If we denote the overall variance as *total*, the within-category variance as *within* (or *unexplained*), and the between-group variance as *between* (or *explained*), then the fundamental ANOVA relationship is

Total variance = Within variance + Between variance.

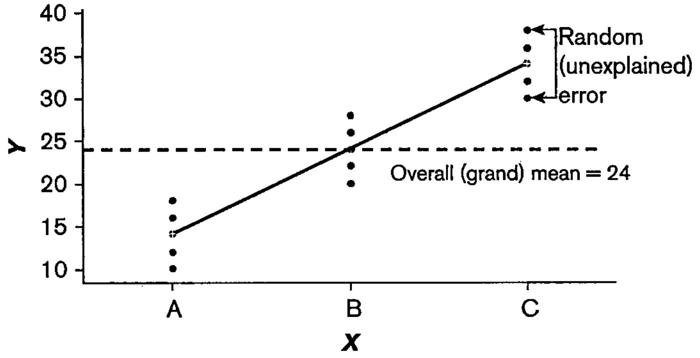
The terms *between* or *explained* refer to the fact that some of the observed differences seem to be due to "membership in" or "having a property of" one category of *X*. That is, on average, the As differ from the Bs. Knowing that a case has characteristic A tells us roughly what its value on *Y* will be. The prediction will not be perfect, however, because of the internal variation among the As. Yet if we could

TABLE 13-24 **Measurements on *Y* within Three Categories of *X***

	Categorical Variable (X)		
	A	B	C
	10	20	30
	12	22	32
	14	24	34
	16	26	36
	18	28	38
Number of cases (N_i)	5	5	5
Mean ($\bar{Y}_i$)	14	24	34
Standard deviation ($\hat{\sigma}_i$)	3.16	3.16	3.16

Overall "grand" mean = 24.

FIGURE 13-7 **Dot Chart of *Y* by Categories of *X*: Means Differ**



Note: Variation in *Y* is due to differences in *X* and "error."

numerically measure these different sources of variability, we could determine what percentage of the total was explained:

Percent explained = (Between/Total) × 100.

HELPFUL HINTS

Variation in ANOVA

TSS stands for the total variability in the data (read this as "the total sum of squares"); *BSS* represents the between means variability ("between sum of squares"); and *WSS* ("within sum of squares") is the within groups or categories variation. With these definitions we have

$$TSS = BSS + WSS, \text{ and}$$

$$\text{Percent explained} = (BSS/TSS) \times 100$$

Check out more Helpful Hints at **edge.sagepub.com/johnson8e**

If the portion of total variance explained by the independent variable is relatively large, there is reason to believe that at least two of the population means are not equal. You can figure out which are different only by examining graphs and calculating effect sizes.

Now look at figure 13-8. It shows two things: the means of A, B, and C are all the same, and the observations differ among themselves but not because they belong to one or another group. Each level of *X* has the same mean. So the total variation in scores has nothing to do with levels of the factor. There is no difference among means and hence no explained or between variation. The analysis of the relationship is thus total variation = within variation + zero, and the percent explained variation is zero:

$$\text{Percent explained} = \text{Explained/Total} = 0/\text{total} = 0.$$

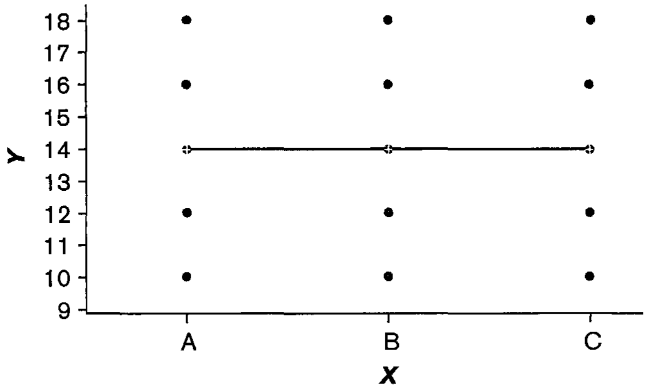
The mathematics of ANOVA simply involves quantifying the types of variation and using these numbers to make inferences.

The percent of variation explained is a commonly used (and abused!) measure of the strength of a relationship. In the context of ANOVA, the percent variation explained is sometimes called **eta-squared** (η^2). One of the properties of eta-squared may be obvious: it varies between zero, which means the independent variable (statistically speaking) explains nothing about *Y*, to 1, which means it accounts for all of the variation. You frequently read something to the effect that "X explains 60 percent of the variation in Y and hence is an important explanatory factor." Whether or not the data justify a statement of this sort depends on a host of considerations, which we take up later.

DO MEANS DIFFER? SIGNIFICANCE TEST FOR ANALYSIS OF VARIANCE. A test of the hypothesis that K subpopulation means are equal (H_0: $\mu_1 = \mu_2 = \mu_3 = \ldots = \mu_K$) rests on several assumptions, especially that the observations in one group are independent of those in the other groups. In addition, we assume large N_Ks and equal population variances (that is, $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \ldots \sigma_k^2$). Test results are most often organized and summarized in an ANOVA table like table 13-25.

The terms inside the table may seem intimidating at first, but the numbers are straightforward. The sums of squares are calculated from formulas described elsewhere. Each sum of squares has an associated degrees of freedom, *df*. They are easy to calculate: the between *df* is the number of categories (K) of the independent variable minus 1, or $K - 1$; and the within *df* is N, the total sample size, minus the number of categories, or $N - K$. Together they sum to the degrees of freedom for the total sum of squares, or $(K - 1) + (N - K) = N - 1$.

FIGURE 13-8 Dot Chart of *Y* by Categories of *X*: Means Do Not Differ



Note: *X* does not "explain" any variation in *Y*.

TABLE 13-25 Typical ANOVA Table Format

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	Observed *F*
Between (name of variable)	*BSS*	$df_{\text{between}} = K - 1$	$\text{BMS} = \text{BSS}/df_{\text{between}}$	$F_{obs} = \frac{BSS/(K-1)}{WSS/(N-K)} = \frac{BMS}{WNS}$
Within (unexplained or error)	*WSS*	$df_{\text{within}} = N - \text{K}$	$\text{WMS} = \text{WSS}/df_{\text{within}}$	
Total	*TSS*	$df_{\text{total}} = N - 1$		

Note: Assume *X* has *K* categories.

Whenever a sum of squares is divided by its degrees of freedom, the quotient is called a mean square. The between mean square is divided by the within mean square to obtain an observed test statistic called the *F* statistic. Symbolically,

$$F_{\text{obs},df_{\text{Numerator}},df_{\text{Denominator}}} = \frac{\dfrac{SS\text{ between}}{df\text{ for between}}}{\dfrac{SS\text{ error}}{df\text{ for error}}} = \frac{\text{Mean square between}}{\text{Mean square error}}.$$

Like the other statistics we have discussed (for example, chi square, *t*, and *z*), the observed *F* has a distribution called (sensibly enough) the *F* distribution. As in the case of the *t* distribution, the *F* distribution is a family, each member of which depends on *two* degrees of freedom, one for the between component and one for the within component. A decision about the null hypothesis of equal population means is made by comparing F_{obs} to F_{crit}.

The general idea should be familiar by now. Suppose we use the hypothetical data in table 13-24 to test the hypothesis that $\mu_A = \mu_B = \mu_C$ against the alternative that at least two of them differ. (Technically, we should have larger samples, but this is just an illustration.) For this test, we choose the .001 level of significance. Table 13-26 shows the results as they are typically spewed out of a computer.

The observed *F* is 50, considerably larger than the critical *F* (with 2 and 12 *df* at the .001 level) of 12.97. (The critical value is found in appendix D.) First decide on a level of significance, then on the degrees of freedom for the within sum of squares (12) and for the between sum of squares (2). The needed value will be in the third *F* table. Since F_{obs} exceeds F_{crit}, the null hypothesis is rejected at the .001 level.

TABLE 13-26 **ANOVA Results**

Source of Variation	Degrees of Freedom (df)	Sum of Squares	Mean Squares	Observed F
Between X	2	1,000	500.0	50.0
Within	12	120	10.0	
Total	14	1,120		
p-level ≈.000				

Source: Table 13-24.

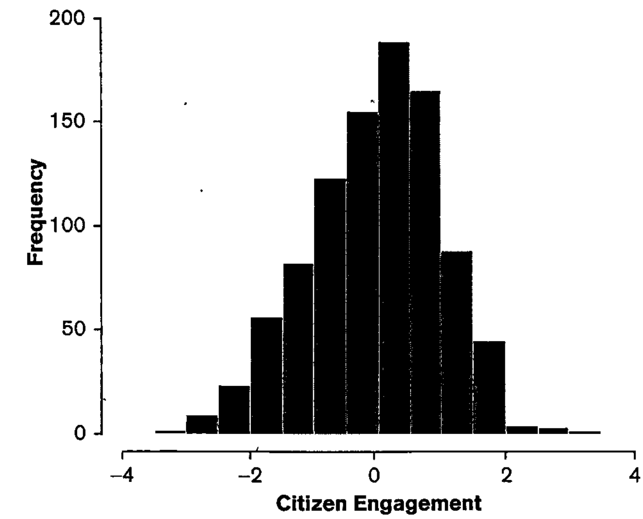
Indeed, if the null hypothesis were in fact true, the p-value (.000) tells us we have obtained a very improbable result.

What does all this mean? So far, we only know that two or more population means are probably unequal. Without looking at confidence intervals, we do not know which differ or by how much. The precise source of the explained variation is not obvious. Once again, at the risk of beating a dead horse, we emphasize that a significance test provides helpful information but does not relieve you of the duty to scrutinize your data from several angles.

For a "real" example, we return to the question of what motivates political participation. The Citizenship, Involvement, Democracy Survey that we've been using contains an interesting item, "Citizenship Norms," which measures "the public's adherence to different potential citizenship norms." To operationalize the concept, the investigators combined responses to a series of questions—"To be a good citizen, how important is it for a person to be . . . [list items]. 0 is extremely unimportant and 10 is extremely important"[24]—into "factor" scales. Although the scale scores (e.g., 1.0) have no intrinsic meaning, they nevertheless allow us to compare respondents on their degree of commitment to *active* citizenship (e.g., voting, being involved in politics and voluntary groups, forming and expressing opinions) as opposed to "obedience" norms (e.g., "always obeying the law," serving in the military, paying taxes). The higher the score, the more a person believes citizenship entails active participation, not mere acquiescence to rules. The summary statistics for this variable are N = 944, mean = –0.036, median = .057 and standard deviation = 1.02, and IQR = 1.387. From these data and the histogram (figure 13-9), we

[24] Marc M. Howard, James L. Gibson, and Dietlind Stolle, "United States Citizenship, Involvement, Democracy (CID) Survey" (Ann Arbor, Mich.: Inter-university Consortium for Political and Social Research, 2007), 159.

FIGURE 13-9 **Engaged Citizen Scores**



Source: Citizenship, Involvement, Democracy Survey, 2006.

see that the sample values seem to be roughly normally distributed with a mean of nearly zero.[25] Note that the means nearly equal the medians and the variation of Y is more or less the same across categories.[26] (See Figure 13-10.)

But what explains the variation in these scores? Are they associated with other factors that might make these data more intelligible? In particular, how will the average scale scores vary among different groups? Right off the bat we might hypothesize that highly partisan citizens (i.e., people who closely identify with one of the two major American political parties) will tend to believe that good citizenship involves more than passively following the rules and the obligations of a democracy include articulating and acting on one's opinions. Nonpartisans, by contrast, will stress nonactivist norms.[27] In short, we can propose that mean engagement scores will increase across categories of partisanship. The null and alternative hypotheses are these:

- H_0: $\mu_{\text{Nonpartisan}} = \mu_{\text{Weak}} = \mu_{\text{Moderate}} = \mu_{\text{Strong}}$.
- H_A: at least two means (μs) are not equal.

[25] As a refresher of chapter 11, you could test the hypothesis that the population mean of engagement is zero. What test would you use?

[26] In technical terms, this is homoscedasticity (equal variances).

[27] For a classic discussion of the definition and analysis of *subject* and *citizen* norms, see Gabriel A. Almond and Sidney Verba's *The Civic Culture: Political Attitudes and Democracy in Five Nations* (Princeton, N.J.: Princeton University Press, 1963).

HOW IT'S DONE

Calculating Sums of Squares[1]

Let N = total sample size and X be the independent variable with K categories: $k = 1, 2, \ldots K$.

First, get totals, T_k, for *each* group or subpopulation:

$$T_k = \sum_{k=1}^{N_k} Y_{ik}, k = 1,2,\ldots K.$$

Calculate three quantities:

- Square each subtotal (T_k), divide by N_k, the total number of cases in the kth subpopulation.

$$A = \sum_{k=1}^{K} \frac{T_k^2}{N_k}$$

- Sum all observations, square the result, the divide total by N:

$$B = \frac{\left(\sum_{i=1}^{N} Y_i\right)^2}{N}.$$

- Square each observation and obtain the total:

$$C = \sum_{i=1}^{N} Y_i^2.$$

Finally, the sum of squares are:

- Total sum of squares:

$$\text{TSS} = \text{C} - \text{B}.$$

- Between sum of squares:

$$\text{BSS} = \text{A} - \text{B}.$$

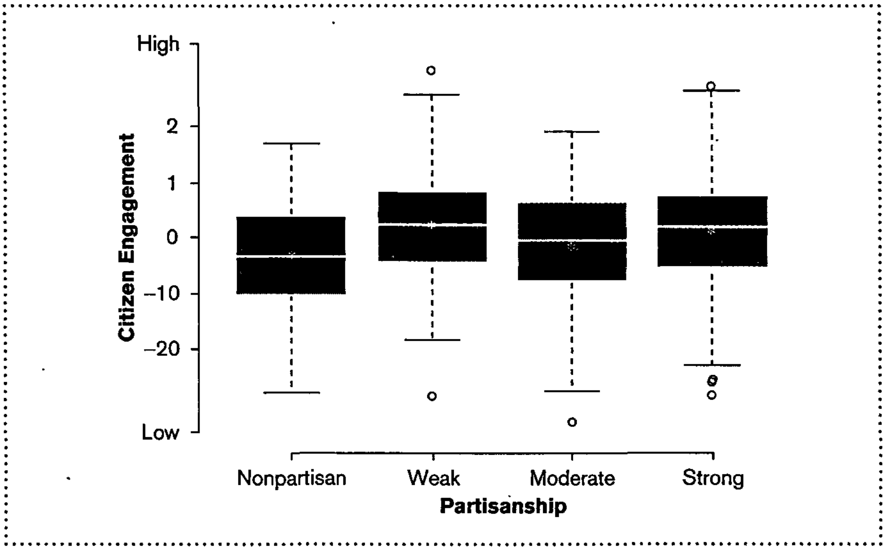
- Within sum of squares:

$$\text{WSS} = \text{C} - \text{A}.$$

Note: If you have a statistical calculator the sum of squares used here and elsewhere can be found by applying the calculating formulas provided below. Most calculators will automatically accumulate both the sum of a set of numbers $\sum_{i=1}^{N} Y_i$ and the sum of their squares $\sum_{i=1}^{n} Y_i^2$.

[1] Based on notes prepared by Richard Williams, University of Notre Dame. Available at Stats I – http://www.nd.edu/~rwilliam/stats1/

FIGURE 13-10 Engagement Scores by Levels of Partisanship



Source: Citizenship, Involvement, Democracy Survey, 2006.

Figure 13-10 offers a bird's-eye view of the data. From the picture it is apparent that the means vary only roughly in the predicted manner and most means differ only marginally. We immediately have reason to doubt that this proposition is going to hold water.

Still, let's proceed with a formal *F*-test. Even though these are survey data and we can use regression analysis to accomplish the same task (see the following discussion and chapter 14), these data lend themselves to analysis of variance if we think metaphorically: consider the levels of partisanship as "treatments" or factors that are "assigned" to individuals whose responses are then recorded. Our job is to compare mean squares: the "explained" (by partisanship) and the "unexplained." Table 13-27 shows the results.

We now confront an interesting situation. The boxplot of the category suggests little real difference in average values across levels of partisanship. Yet the *F* statistic (9.98) is highly significant, which tells us that after rejecting the null hypothesis, we should search for which means differ from which in meaningful ways. So which ones differ? After all, there six comparisons: "strong" versus "nonpartisan," "strong" versus "weak,". . . "moderate" versus "weak." Are all subpopulations different, or only a subset of them? Do any differences have a substantive meaning? A picture might clarify matters.

TABLE 13-27 **ANOVA Table and *F*-Test**

Source of Variation	Degrees of Freedom (*df*)	Sum of Squares	Mean Squares	Observed F
Between (due to partisanship)	3	30.22	10.07	9.98*
Within	940	948.22	1.02	
Total	943	978.44		
*p-level ≈.000				

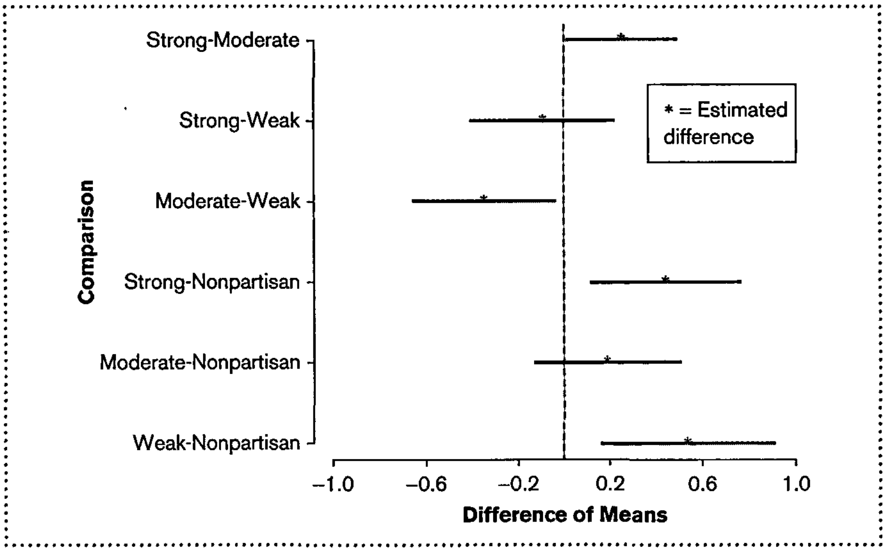
Source: Citizen, Involvement, Democracy Survey, 2006.

Carefully examine figure 13-11. The stars (*) represent the six estimated differences. The bars around them are 99 percent confidence intervals.[28] These intervals contain a range of values that probably includes the true differences. Look, for instance, at the strong-nonpartisan interval shown by third bar from the bottom. It does not include zero as a likely value. So we might conclude that strong partisans in the population have on average higher engagement scores than nonpartisans. (We know the value is "greater" because the difference is positive.) Similarly, the next bar indicates that moderates and nonpartisans do not differ significantly because the line includes zero, suggesting that the difference could well be zero. As you go through the graph, you can see that at least two and possibly three pairs of means differ. But apart from these there does not seem to be a clear pattern in the data. Our expectation at the outset was that as partisanship increased, so too would citizen engagement scores. Yet, as figures 13-11 and 13-12 reveal, there is no obvious trend in the data.

We should also compare the magnitudes of the differences. Doing so is difficult, however, because the scale scores are abstract numbers that show respondents' positions on a scale constructed from ten individual items. Thus, the difference between strong partisans and independents is 0.44. It's statistically significant, but is it meaningful? In the absence of an understandable scale, we have to

[28] An important technical issue arises in the calculation of these intervals. As we said when discussing difference-of-means tests, it is usually not advisable to compute a series of *t*- or *z*-tests to see which means differ. The reason is that probability statements—"the difference between group *X* and group *Y* is significant at the .01 level"—will be incorrect unless all tests are truly independent of one another. With more than two comparisons on the same data, this requirement is not met. Hence, statisticians use "simultaneous inference" to construct tests and confidence intervals. The particular application used here is called "Tukey's 'Honest Significant Difference'" method after John Tukey, a preeminent statistician at Bell Labs. See Brian S. Yandell, *Practical Data Analysis for Designed Experiments* (New York: Chapman and Hall/CRC, 2007).

FIGURE 13-11 **Confidence Intervals for All Differences among Means**



Source: Citizenship, Involvement, Democracy Survey, 2006.

rely on patterns and trends as shown by graphs for the most informative interpretation.

In any case, the results demonstrate the limitations of a hypothesis test and the necessity of using a variety of graphs as well as statistics to construct interpretative pictures.

Regression Analysis

The procedure, **regression analysis**, is really a toolbox of methods for describing how, how strongly, and under what conditions an independent and dependent variable are associated. The regression framework is without doubt the meat and potatoes of empirical social and political research because of its versatility. Regression techniques can be used to analyze all sorts of variables having all kinds of relationships. But its value goes even further. In chapter 6 we discussed the difficulty of making causal inferences on the basis of nonexperimental observations. Many scholars believe regression analysis and related approaches act as reasonable surrogates for controlled experiments and can, if applied carefully, lead to causal inferences. As with many other techniques, the first step is to get the overall picture.

HELPFUL HINTS

A Reminder about Confidence Intervals and Tests of Statistical Significance

In this text, we describe both hypothesis tests and confidence intervals. The former are common in scholarly and popular reports of statistical findings, but the latter, we believe, give you all the information a hypothesis test does, and then some.

Suppose you hypothesize that in a population, $A - B = 0$. You conduct a *t*- or *z*-test on a sample and reject this hypothesis at the .05 level. This means you are pretty sure A does not equal B. You are not positive, of course, but if you carried out the analysis correctly, there is only a 1 in 20 (.05) chance of rejecting a hypothesis that should not be rejected.

What about finding 95 percent confidence limits for the $A - B$ difference? Notice that $1 - .95 = .05$. The connection between .95 and .05 is not coincidental. A confidence interval puts a positive spin on your inference: you are 95 percent sure that the true difference lies somewhere in the interval. By contrast, a significance test at the .05 level is in a sense a negative statement: you think there's only a 5 percent chance (.05 probability) that you are wrong. But as far as probability theory is concerned, one procedure is as good as another.

So why do we prefer confidence intervals? Because besides allowing you to discard (or not) a hypothesis, confidence intervals explicitly show you possible values of the effect; better still, they are presented in the measurements of the variables in the problem at hand. If you reject the hypothesis that A equals B, you only know that B probably does not equal B. But if the confidence intervals for the difference run from, say, 1 to 2, you may conclude that there is not a meaningful theoretical or practical difference. On the other hand, if the interval extends from 1 to 50, that might suggest something worthy of further investigation.

Check out more Helpful Hints at **edge.sagepub.com/johnson8e**

Basic Assumptions

A linear regression model for two variables has this form:

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon.$$

The betas are regression parameters (constant and slope), and ε_i is the error terms. Before undertaking estimation and significance tests for the regression parameters, we need touch on a topic that we have so far finessed: Under what conditions is it reasonable to develop, estimate, and test linear models? That is, since we only have samples or actual realizations of data, we can only know so much about the population from which they came and the causal mechanisms at work. Much of the rest we have to assume. The assumptions, some of which can be verified to one degree or another, include the following:

- Correct specification: The model includes all the necessary independent variables and excludes the unnecessary ones. By *necessary,* we mean the variables that systematically affect Y. Unnecessary information in the form of irrelevant variables just increases errors of prediction.
- Linearity: The expected value of the dependent variable is a *linear* function of the independent variable. Thus, if X_1 is an explanatory factor, its impact on Y comes in the form of $\beta_0 + \beta_1 X$—not, say, exponentially, as in β_1^X. (It is possible, however, to create new variables and use them additively. Hence, we could define $Z = \beta^X$. But of course now a unit change in Z will be harder to interpret.)
- X measured without error: A critical assumption is that the independent variable is measured without systematic error. To take a quick example, suppose we have observed X values but in reality our measurement instrument is faulty, so the observed Xs are a function of "true" values and an error: $X = x$ + error, where x is the actual value. In this situation, regression estimates may be biased, which gives us a reason to think carefully about the independent variables.
- Independent observations: Data are collected in such a fashion that the inclusion of one case has no bearing on the selection of any other (as opposed to, say, sampling ten individuals and then including their best friends).
- Random errors: The error component, which represents the effects of omitted causes of Y, measurement errors *in* Y, and "natural" variation among subjects, must be truly random in the sense that the errors cancel. In statistical language, $E(\varepsilon_i = 0)$, which is read as "the expected (long-range) value of the errors is zero."
- Constant error variance: The variation in errors is the same for each level or value of X. (Want a fancier term? Constant variance is called *homoscedasticity* or *homogeneity* of variance.)
- Normally distributed errors: You can think of the errors added onto the linear model as unseen, unmeasured variables. Nevertheless, they still have a distribution, which (for testing hypotheses) we assume is normal with a mean of zero and variance σ^2.

This list serves an important purpose: to remind us that a modeling technique like linear regression carries a lot of hidden baggage. Just because an analyst ignores them does not mean the effects of violating assumptions go away. A computer can crank out reams of respectable-looking reports without the results being accurate or meaningful. Consequently, there is a vast and continuously growing statistical literature on how bad the violations are and what to do about them. Unfortunately, the topic is too large and complex to explore here. Thus, we generally proceed as if the assumptions were true.[29]

Scatterplots

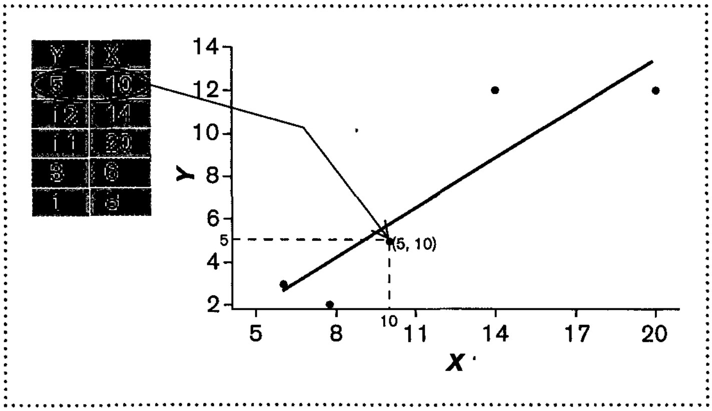
The beginning point of regression analysis is the identification of associations or correlations between pairs of variables, and graphs provide the best first step.

One common graph is the scatterplot. Intended for quantitative data, a **scatterplot** contains a horizontal axis representing one variable and a vertical axis (drawn at a right angle) for the other variable. The usual practice is to place the values of the independent variable along the *x*-axis and the values of the dependent variable along the *y*-axis. The scales of the axes are in units of the particular variables, such as percentages or thousands of dollars. The *X* and *Y* values for each observation are plotted using this coordinate system. The measurements for each case are placed at the point on the graph corresponding to the case's values on the variables.

As an example, figure 13-12 shows five *Y* and *X* values and how they are plotted on a scatterplot. Each case is located or marked at the intersection of the line extending from the *x*- and *y*-axes. The first pair of observations, 5 and 10, appears at the point $Y = 5$ and $X = 10$.

Scatterplots are handy because they show at a glance the form and strength of relationships. In this example, increases in *X* tend to be associated with increases in *Y*. Indeed, we have drawn a straight line on the graph in such a fashion that most observations fall on it or very near to it. In the language introduced at the chapter's outset, this pattern of points indicates a strong "positive linear correlation."

FIGURE 13-12 **Construction of a Scatterplot**



Y	X
5	10
12	14
11	20
3	6
1	8

Source: Hypothetical data.

[29] A thorough treatment of the assumptions underlying regression analysis is John Fox's *Applied Regression Analysis and Generalized Linear Models*, 2nd ed. (Thousand Oaks, Calif.: SAGE, 2008), chap. 6.

HELPFUL HINTS

A Tip and a Warning

If you have a large batch of quantitative data (say, more than 500 cases), you can obtain clearer, more interpretable results if you ask your software to first select a *sample* of the data (25 to 75 cases) and plot those numbers. If the sample is truly representative, the plot will reveal the important features of the relationship. Creating a scatterplot from an entire dataset may produce a picture filled with so many dots that nothing intelligible can be detected. Furthermore, scatterplots are suitable only for quantitative variables; they are not intended for categorical (nominal and ordinal) data. If you tried, for instance, to get your software to plot party identification by gender, the result would be two parallel lines that would tell you nothing.

Check out more Helpful Hints at **edge.sagepub.com/johnson8e**

For a more realistic illustration, let's return to the discussion of the causes and consequences of inequality. Look at figure 13-13. It shows how union density (percentage of labor force who are members of unions) is related to a measure of income inequality (Gini scores) in twenty-one developed countries. The data, which come from table 11-1, show each nation's values on union density (X) and Gini scale (Y). We have also added a so-called least-squares line to underscore the decreasing, approximately linear pattern of the relationship. Finally, we tagged two countries, Japan and the United States, to illustrate how the graph is constructed. (If you refer to table 11-1, you will see that the union–Gini values for Japan are 20.3 and 24.9, respectively; those for the United States are 12.6 and 40.8.) We see a familiar pattern in the scatter of the points: correlation. It has these features:

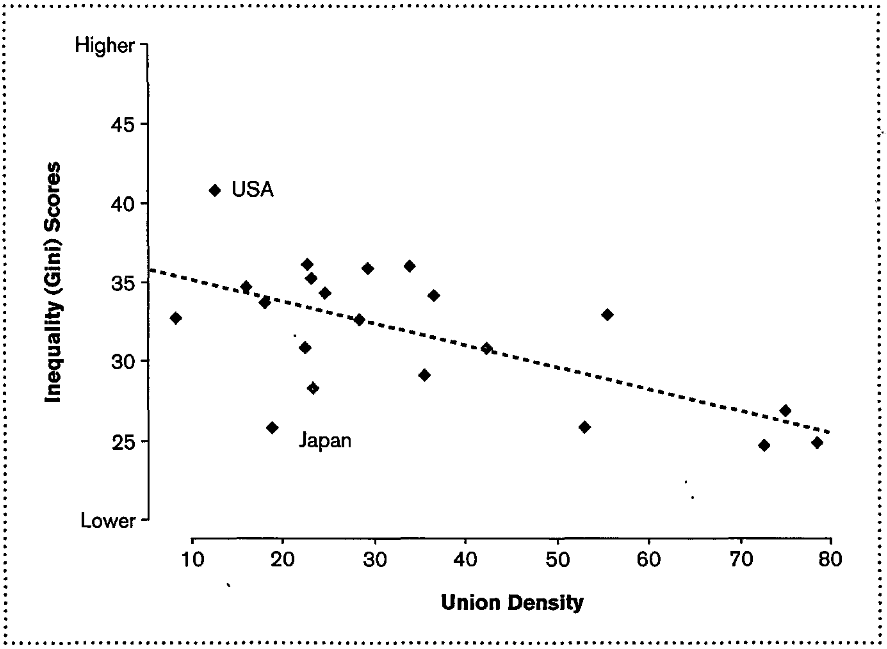
- The association is roughly linear; the points lie near a straight line.
- The correlation is negative; the line slants downward, telling us that an increase in unionization is associated with a decrease in inequality. Bluntly stated, the more unions, the less the inequality.
- The correlation is moderately strong: the points don't form a perfect linear pattern, but the configuration is clear enough.

- The two identified cases (Japan and the United States) provide further examples of how to interpret the plot. If the straight line is used to predict or estimate a country's Gini coefficient on the basis of union density, we see that Japan has a lower than expected score, while the US score is higher. These are similar labor profiles but quite different outcomes. (But in one sense, the errors almost cancel—a point we discuss in greater detail later.) Most of the observations lie nearer the line, which suggests that whatever technique produced the line might provide a method to quantify the correlation more precisely.

The Linear Regression Model

The examination of scatterplots is a good first step in describing statistically or modeling a two-variable relationship. Figure 13-13 shows the relationship between union density and a measure of income inequality, the Gini variable, in twenty-one developing nations. It reveals a pattern: *large* values of density are associated with *small* values of inequality, and vice versa. The relationship is by no means perfect, but there does seem to be a negative correlation.

FIGURE 13-13 **Inequality by Union Density**



Source: Table 11-1.

As mentioned earlier, we added a line to show the correlation. The slope of this line is a negative number, which means that as we go up the scale on the x-axis, we move down the y-axis. These ideas can be further clarified by recalling high school algebra. The equation for the graph of a straight line has the general form

$$Y = a + bX.$$

In the equation, X and Y are variables. The first letter is called the constant and equals the value of Y when X equals zero (just substitute 0 for X). Also, the constant equals the mean of the dependent variable or $a = \overline{Y}$, a fact that turns out to be useful.

The equation for the linear model has a geometric interpretation as well. If the graph of the equation is plotted, a is the point where the line crosses the y-axis. The letter b stands for the slope of the line, which indicates how much Y changes for each one-unit increase in X. For a positive b (i.e., $b > 0$), if we move up the X scale one unit, b indicates how much Y increases. (If we applied this type of equation to the data shown in figure 13-13, b would be negative and would indicate how much Y, Gini scores, decline for every unit increase in X, union concentration.) If there is no (linear) relationship between the two variables, the slope of the line is zero, and its graph is horizontal and parallel to the x-axis.

Note that the line's slope depends partly on the measurement scale of X. So, if one were inadvertently to use Y as the dependent variable, a slope could be calculated, but its magnitude would in general *not* be the same as if X were treated as independent. Recall from the beginning of the chapter that in the language of statistics, the slope is an *asymmetric* parameter.

Regression Basics

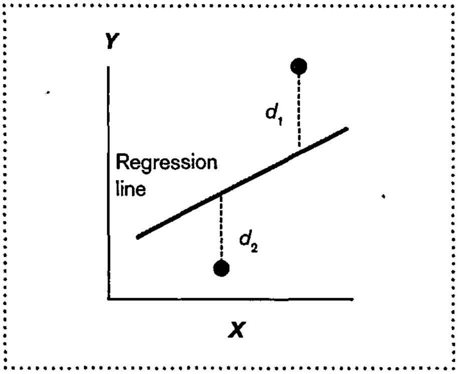
Regression analysis applies these ideas to two variables, where both are numeric or quantitative. (Actually, regression analysis is general enough to include categorical variables, but only in special ways. We discuss this possibility in chapter 14.) The goal here is to find the constant and slope of an equation that "best fits" the data.[30]

What exactly does "fit" mean in this context? In regression, an equation is found in such a way that its graph is a line that minimizes the squared vertical distances between the data points and the line drawn. In figure 13-14, for example, d_1 and d_2 represent the distances of observed data points from an estimated regression line.

[30] Consider, for example, a model that contains two types of variables—one group measuring demographic factors and another measuring attitudes and beliefs. The investigator might want to know if the demographic variables can be dropped without significant loss of information.

This mathematical procedure is called least squares and is often called "ordinary least squares," or OLS for short.

FIGURE 13-14 **Data Points Do Not Fall on Regression Line**



As noted several times before, we utilize lowercase Greek letters to denote unknown quantities and Greek letters with hats (^) over them to denote estimators of these numbers. In the two-variable case, the regression model commonly appears as

$$Y_i = \beta_0 + \beta_{YX}X + \varepsilon_i.$$

Pay attention to the subscripts on the second beta. They signify that we are regressing "Y on X," not the other way around. That is, the dependent variable is Y and is listed first in the subscript. The independent variable, X, comes second. *Always remember this key point: regression analysis is asymmetric in that the choice of dependent variable matters because, as noted in the previous section, the numerical value of the slope (regression coefficient) depends on which variable is considered dependent.* Throughout we have treated Y as the dependent variable. Were we to switch their roles, we would be regressing X *on* Y. And usually $\beta_{YX} \neq \beta_{XY}$.

The regression model contains two parameters to estimate and test. The first, β_0, is the constant and is interpreted exactly as indicated before: it is the value of Y when X equals zero. (Remember that in the simple case, it also equals the mean of Y.) The second, β_1, is the **regression coefficient** and tells how much Y changes per unit change in X. The regression coefficient is always measured in units of the dependent variable.

These quantities are calculated by the formula:

$$\hat{\beta}_{YX} = \frac{N\sum_{i=1}^{N}X_iY_i - \left(\sum_{i=1}^{N}X_i\right)\left(\sum_{i=1}^{N}Y_i\right)}{N\sum_{i=1}^{N}X_i^2 - \left(\sum_{i=1}^{N}X_i\right)^2},$$

and

The regression constant is: $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_{YX}\bar{X}$,

where $\bar{Y}$ and $\bar{X}$ are the means of Y and X, respectively, and $\hat{\beta}_{YX}$ is the regression coefficient as calculated above.

The error (ε) indicates that observed data do not follow a neat pattern that can be summarized with a straight line. It suggests instead that an observation's score on Y

can be broken into two parts: one that is "due to" the independent variable and is represented by the linear part of the equation, $\beta_0 + \beta_1 X$, and another that is "due to" error or chance, ε. In other words, if we know an observation's score on X and also know the equation that describes the relationship, we can substitute the number into the equation to obtain a *predicted* value of Y. This predicted value will differ from the observed value by the error:

Observed value = Predicted value + Error.

If there are few errors—that is, if all the data lie near the regression line—then the predicted and observed values will be very close. In that case, we would say the equation adequately explains, or fits, the data. In contrast, if the observed data differ from the predicted values, then there will be considerable error, and the fit will not be as good.

Figure 13-15 ties these ideas together. Suppose we consider a particular case. Its scores on X and Y (X_i and Y_i) are represented by a dot (•). Its score on X is denoted as X_i. If we draw a line straight up from X_i to the regression line and then draw another line to the y-axis, we find the point that represents the predicted value of Y, denoted $\hat{Y}_i$. The difference between the predicted value, $\hat{Y}_i$, and the observed value, Y_i, is called the error or **residual**. A residual represents the difference between a predicted score based on the regression equation—which is the mathematical equation describing the relationship between the variables—and the observed score, Y_i. (As we see in a moment, it stands for that part of a Y score that is unexplained.) Regression-computing formulas pick values of α and β that minimize the sum of all these squared errors.

HOW IT'S DONE

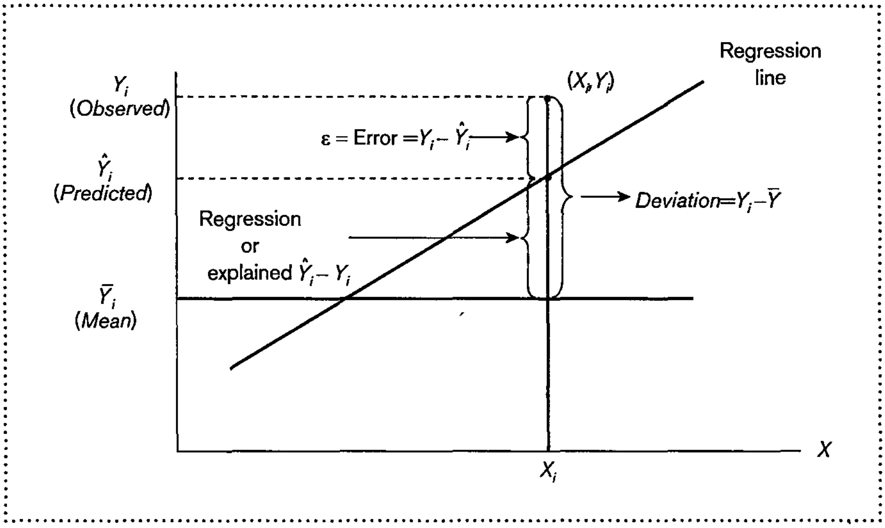
Calculation of Estimated Regression Coefficients

The regression coefficient is calculated as follows:

$$\hat{\beta}_{YX} = \frac{N\sum_{i=1}^{N} X_i Y_i - \left(\sum_{i=1}^{N} X_i\right)\left(\sum_{i=1}^{N} Y_i\right)}{N\sum_{i=1}^{N} X_i^2 - \left(\sum_{i=1}^{N} X_i\right)^2},$$

where N is the number of cases and X_i and Y_i are the X–Y values of the ith case. The regression constant is calculated as follows: $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_{YX}\bar{X}$, where $\bar{Y}$ and $\bar{X}$ are the means of Y and X, respectively, and $\hat{\beta}_{YX}$ is the regression coefficient as calculated.

FIGURE 13-15 Predicted and Observed Values of Y



Although the minimizing procedure may sound complicated, computer software and many handheld calculators make finding regression equations relatively easy. The tricky part is understanding the results. We do not show the calculations here, but the estimated equation for the regression of Gini scores on union density is

$$\hat{Y} = 36.44 - .14\text{union}.$$

What exactly do the numbers 36.44 and −.14 mean?

The slope of the line and the *y*-intercept describe the nature and strength of the connection between the two variables. Again, the *y*-intercept is the value of the dependent variable when X (the independent variable) = 0, or, stated differently, it is the place where the regression line crosses the *y*-axis when $X = 0$.

In the current example the *y*-intercept of 36.44 means that when a country's unionization is zero, its predicted Gini score is about 36.44. This value is, of course, a prediction for a country with no unions. In many instances, the actual value is not of much substantive interest because a zero value on the independent variable does not make much theoretical sense.

The slope or regression coefficient is a different matter, however. It measures the amount the dependent variable changes when the independent variable, X, changes one unit. In this case, the slope of −.14 tells us that for every 1 percent increase in union strength, there is a predicted or estimated *decrease* of .14 units of "inequality." (We label the variables "union strength" and "inequality" to keep reminding us that they refer to politically interesting and important concepts. But the actual measurements—the operational definitions of strength and inequality—were described earlier.) The two variables appear to be linked in the hypothesized way. Remember, the initial substantive problem was to explain cross-national variation in inequality. The literature we cited earlier claims that the large disparities we see between the richest, say, 1 percent and the rest of the population do not occur by chance but result from political struggle. Our idea is that the better equipped the lower classes are to press their demands, the greater share of the national wealth they can obtain for themselves. When looked at from the narrow point of view of our research design and dataset, this notion seems to hold water.

The value of the regression coefficient (–.14) may seem small and abstract. Here's a simple method to help you better understand its meaning. Try performing some thought experiments in which you systematically substitute "informative" values of X into the equation to observe how they affect the dependent variable. Let's take some arbitrary but practically meaningful union density scores and insert them one by one into the estimated equation, $\hat{Y}_i = 36.44 - .14$ union, to see their effects via predicted values. Table 13-28 shows the outcome. The entries are obtained by substitution as illustrated here:

$$\begin{aligned}\hat{Y}_i &= 36.44 - .14(0) = 36.44\\ \hat{Y}_i &= 36.44 - .14(1) = 36.30\\ \hat{Y}_i &= 36.44 - .14(10) = 35.04\\ &\ldots\end{aligned}$$

The predicted Gini score of a country without any organized labor (0 percent) is 36.44. Suppose its union rate increased 1 percentage point. Line 3 in the table tells us that its score would drop to 36.30, a decline of .14 units. This amount is exactly the estimated regression coefficient reported above, confirming intuitively that it measures the impact of a one-unit change in X. Now, is the effect, .14, a big deal or not? It bears repeating that the regression parameter is measured in units of the dependent variable, here income inequality. In practical terms, a 1 percent change in union participation would not much affect a country's labor relations profile. But what if it increased 10 percent (or 10 units)? The predicted Y is 35.04. When rounded, the decrease is 1.4, which is roughly 10 times the slope, as it should be.[31] Or, if half a nation's eligible workers belong to unions, the predicted score would be 29.44, about a 7-point decline; if all were unionized it would drop all the way to 22.44. (We are implicitly applying a causal interpretation merely to explain how the regression coefficient can be interpreted.)

[31] X was increased 1 × 10 = 10.

TABLE 13-28 **Predicted Gini Scores**

Interpretation	Selected *X* Value (union density)	Predicted *Y* (Gini)
No unions	0	36.44
1 percent unionized	1	36.30
10 percent unionized	10.00	35.04
Minimum observed	8.20	35.29
25th quantile	22.40	33.30
Median	28.20	32.49
Mean	34.79	31.57
75th quantile	42.30	30.52
Observed maximum	78.00	25.52
100 percent unionization	100	22.44

True, it is hard to grasp fully numbers like these when the measurement instrument is this abstract. That's why we later resort to a rescaling of a variable like the Gini coefficient. First, however, consider another short example.

Briefly return to the debate about the existence of a political gender gap. Those who believe that, for whatever reason, women tend to be more liberal than men need to show that the sexes' political attitudes differ to some extent. The 2008 American National Election Study, part of the series of voting studies we've been using, contains variables for gender and political ideology. The latter is a 7-point scale that extends from 1 (*most liberal–least conservative*) to 7 (*least liberal–most conservative*). For the moment, we can treat the scale scores as if they were numbers and use them in a regression analysis as the dependent variable.

But we now have an independent variable, gender. It has been mentioned several times that a dichotomous variable—that is, a variable with two categories—can be given a numerical interpretation by assigning numeric codes to the two categories. It is especially helpful to use codes 0 and 1. This coding scheme is called "dummy." Thus, men are coded 0, women 1. A "one-unit" change in this variable simply means moving from one category to the next. (Again, we are speaking metaphorically because such changes are usually physically impossible or meaningless.) In short, we want to predict ideology based on one's gender. After crunching the numbers, the estimated equation is $\hat{Y} = 4.24 - .18$ (*gender*).

The regression constant in the case of dummy coding has a particularly clear interpretation: it is the mean of Y for the members of the category coded 0. Try putting 0 in the above equation. What do you get? The answer ($\hat{Y} = 4.24 -.18(0) = 4.24$) is the average liberalism–conservatism score among just women. The "effect" of being a male is to reduce the mean ideology score by .18 units: $\hat{Y} = 4.24 - .18(1) = 4.06$. (This number, as you might have guessed, is the mean ideology score for men.) Whether or not this result is worth shouting about remains to be seen. The answer depends on how well the data fit the model.

Remember: the regression parameter is asymmetric. In symbols, $\beta_{YX} \neq \beta_{XY}$, except in certain situations. Had we used Gini rates as the dependent variable (thereby considering union density as independent) we would get a different equation:

$$\hat{Y}_{(\text{union})} = 126.52 - 2.89X_{(\text{Gini})}.$$

This equation is completely different from the previous one. (Of course, the same interpretation applies—"A change of 1 in Gini score is accompanied by (causes) a decrease of about 2.89 percent in unionization."—and in this context leads to the same theoretical conclusion, namely that inequality and union strength are negatively related. But in other situations, mixing the order of variables may lead to nonsensical results, especially if there is a hint of causality in the analysis. Suppose, for example, you were studying the interaction between age and income. If you treated age as the dependent variable and regressed it on income, the resulting regression coefficient could be interpreted "as a one dollar increase in income leads to a specified amount of aging"—clearly a silly conclusion. The moral is to ponder the choice of dependent and independent variables.

Measuring the Fit of a Regression Line: R^2

Let us pause for a moment to glance at figure 13-15 again. Earlier in the chapter, we introduced a term called the *total sum of squares* (*TSS*). It was the sum of squared deviation from the mean. Now, by examining figure 13-15 you can see that an observation's *total* deviation from the mean, denoted $Y_i - \bar{Y}$, can be divided into two additive parts. The first is the difference between the mean and the predicted value of Y. Let's label that portion as the "regression," or "explained," part (*RegSS*). It is explained in the sense that a piece of the deviation from the overall mean is accounted for or attributable to *X*, the independent variable. The second component of the total deviation is called "residual sum of squares" (*ResSS*) and measures prediction errors. This term is frequently labeled the "unexplained sum of squares" because it represents the differences between our predictions—that is, $\hat{Y}$—and what is actually observed. If all the predictions were perfect, there would be no errors, and the residual sum of squares errors would be zero. The residual sum of squares provides the numerator of the "conditional" standard deviation of Y, a statistic used later on to test hypotheses and construct confidence intervals.

These three quantities are identical to the ones presented earlier in connection with the analysis of variance, except two of them have slightly different names—they are now called the "regression" and "error" residual sum of squares. (Their computing formulas also differ slightly.) Yet the same fundamental relationship still holds:

$$TSS = RegSS + ResSS.$$

The total sum of squares (*TSS*) represents all the variation in the data, explained or not, whereas the regression sum of squares (*RegSS*) corresponds to that part of this total that is "explained" (in a statistical sense) by the independent variable via the regression equation. So, as in ANOVA, we can calculate the "proportion of total variation explained by *X*" as

$$R^2 = RegSS/TSS.$$

This measure (R^2) is known as **R-squared** and is one of the most commonly reported statistics in the social sciences.[32] For example, if R^2 is multiplied by 100, the result is often interpreted as the percentage of (total) variation in *Y* that *X* "explains." R^2's popularity derives partly from its simplicity and partly from the belief that it indicates how well a regression model fits data.[33]

Table 13-29 shows the sums of squares from the regression of income inequality on union density. The explained variation is .39, which, statistically speaking, means that *X* explains somewhat less than half of the variation in Y.

An R^2 of .39 means that about 40 percent of the variation in Gini scores is statistically "explained" by union density. This result once again suggests a modest correlation between the two variables. The evidence is consistent with the hypothesis that to the extent that workers are politically empowered, inequality is reduced, but lots more information is needed to support a causal connection.

An important property of *R*-squared is that it is symmetrical, meaning that it has the same value no matter which variable is treated as dependent. This is a key difference from the regression coefficient, which does change depending on the choice of dependent variable. Also, *R*-squared must be at least zero; it cannot be negative because it is the quotient of two squared terms.

[32] *R*-squared is also called the "multiple correlation coefficient," "multiple *R*," and the "coefficient of determination." These terms usually come into play when analyzing the effect of several independent variables.

[33] Computer programs usually report the calculated or obtained probability of the observed chi square, so we do not even have to look up a critical value in a table.

TABLE 13-29 **Regression Sums of Square and *R*-Squared and *r***

Source	Value
Regression (*RegSS*)	157.38
Residual (*ResSS*)	242.68
Total (*TSS*)	400.06
R^2= 157.38/400.06 = .39 (39%); r=-.62	

Figure 13-16 offers some additional insights into the properties and interpretation of *R*-squared. In the first set of graphs (a), we see that if all the data points lie on a straight line, there will be no residual or unexplained deviations, and consequently *X* explains 100 percent of the variation in *Y*. This is true for both perfect positive ($\hat{\beta} < 0$) or negative ($\hat{\beta} < 0$) relationships. Hence, R^2 equals 1. However, if the points have a general tendency to lie on a positively or negatively sloping line, R^2 will be less than 1 but will indicate that some portion of the variation in *Y* can be attributed to *X*. (See section b of figure 13-16.) Finally, if no linear relationship exists between *X* and *Y*, R^2 will be zero. A value of zero means only that there is no relationship describable by a straight line. It does *not* mean statistical independence. The variables may have no association at all, or they may be strongly curvilinearly related or connected in some other fashion (see figure 13-16, section c). In both situations, R^2 will be zero or close to zero, but the meaning will differ. A good way to spot the difference is to examine a plot of *Y* versus *X*. A scatterplot can help you determine the pattern your data come closest to.

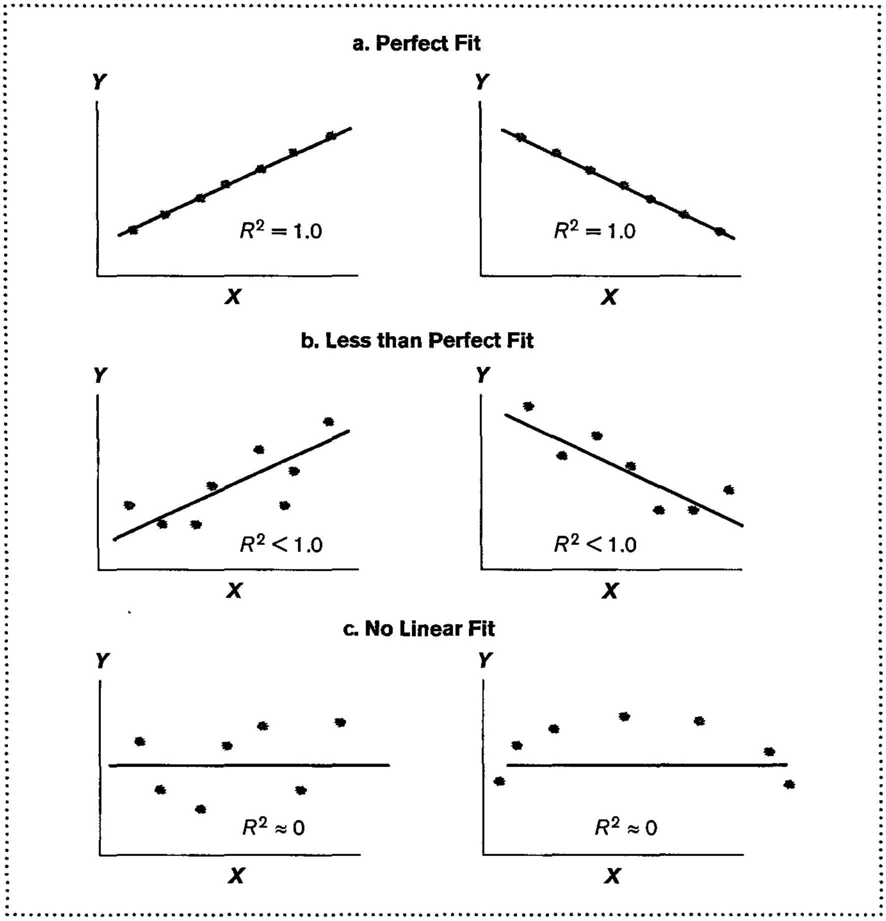
In regression analysis, the term *explained* has a different meaning than it does in day-to-day conversation. In statistics, it means that the variation in one variable can be mathematically divided into two quantities. One, the so-called explained part, is the squared sum of differences between predicted values and the overall mean. These are strictly statistical terms. Thus, we might find a large *R*-squared between two variables (for example, literacy and economic development), which in statistical terms implies that a lot of variation has been explained. But this finding does not necessarily indicate that we really understand why and how countries with higher literacy rates achieve more economic development. In fact, as we explained in chapter 6, a relationship can be spurious, meaning that a false connection is caused by other factors. Always be cautious when confronted with seemingly large values of *R*-squared.

The Correlation Coefficient

A close kin of *R*-squared is the **correlation coefficient**, a measure of the strength and direction of the linear correlation between two quantitative variables. The definition and computation of the correlation coefficient, denoted *r*, depend on standardizing the covariation between *Y* and *X* by dividing by their standard deviations. To find the covariation between two variables, you multiply the "deviations from the mean"—as shown in chapter 12, a deviation is a value minus the mean—and add them together. This total is then divided by the variables' standard deviations.

We listed many general properties of measures of association at the beginning of this chapter, and you may wish to refresh your memory because the correlation

FIGURE 13-16 Values of *R*-Squared



coefficient exhibits many of those properties. Known under a variety of labels—the product-moment correlation, Pearson's *r*, or, most plainly, *r*—this coefficient reveals the direction of a regression line (positive or negative) and how closely observations lie near it. Its properties include the following:

- It is most appropriate if the relationship is approximately linear.
- Its value lies between –1 and 1. The coefficient reaches the lower limit when *Y* and *X* are perfectly negatively correlated, which is to say all the data lie on a straight line sloping downward from left to right. Its maximum value (1) is achieved when the variables are perfectly positively correlated.
- It will equal zero if there is no linear correlation or, to be more exact, when the slope is zero.

- The closer r is to either of its maximum values, the stronger the correlation. The nearer to zero, the weaker the correlation. (Consequently, $r = .8$ or $-.8$ implies a stronger relationship than $r = .2$ or $-.2$.)
- It has the same sign as the regression coefficient. (For example, if $\hat{\beta}$ is negative, r will be, too.)
- Unlike the regression parameter, it is symmetric in that its numerical value does not depend on which variable is considered dependent or independent.
- The correlation coefficient is scale-independent in that its magnitude does not depend on either variable's measurement scale. It does not matter if, say, X is measured in dollars or thousands of dollars; the value r stays the same. This is not true of the regression coefficient.

Because of the last property, the correlation coefficient can be regarded as a kind of regression coefficient that does not depend on the units of Y or X. As a matter of fact, r has this association with the slope:

$$r = \left(\frac{\hat{\sigma}_X}{\hat{\sigma}_Y} \right) \hat{\beta},$$

HOW IT'S DONE

Calculating Sums of Squares and R^2

Let Y_i and X_i be the values for the i^{th} case on Y and X, respectively;

let $\hat{\beta}$ be the regression coefficient from regressing Y on X; let N be the number of cases in the sample; let $\hat{Y}_i$ be the predicted value for ith observation; and let S_X^2 be the sample variance of x (the *independent*) variable found by dividing the total variation in X,

$$\left(\sum_{i=1}^{N} \left(X_i - \bar{X} \right)^2 \right),$$

by N, not $N - 1$.

Total Sum of Squares:

$$TSS = \sum_{i=1}^{N} \left(Y_i - \bar{Y} \right)^2 = \sum_{i=1}^{N} Y_i^2 - \frac{\left[\sum_i Y_i \right]^2}{N}.$$

$$\text{Re}gSS = \sum_{i=1}^{N} \left(\hat{Y}_i - \bar{Y} \right)^2 = \hat{\beta}_{YX} \sum_{i=1}^{N} Y_i X_i = N \hat{\beta}_{YX}^2 s^2{}_X.$$

ResSS = TSS – RegSS

$$R^2 = \frac{\text{Re}\,gSS}{TSS}.$$

HOW IT'S DONE

The Correlation Coefficient

The (Pearson) correlation coefficient is calculated as follows:

$$r = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(X_i - \bar{X})}{\hat{\sigma}_Y \hat{\sigma}_X},$$

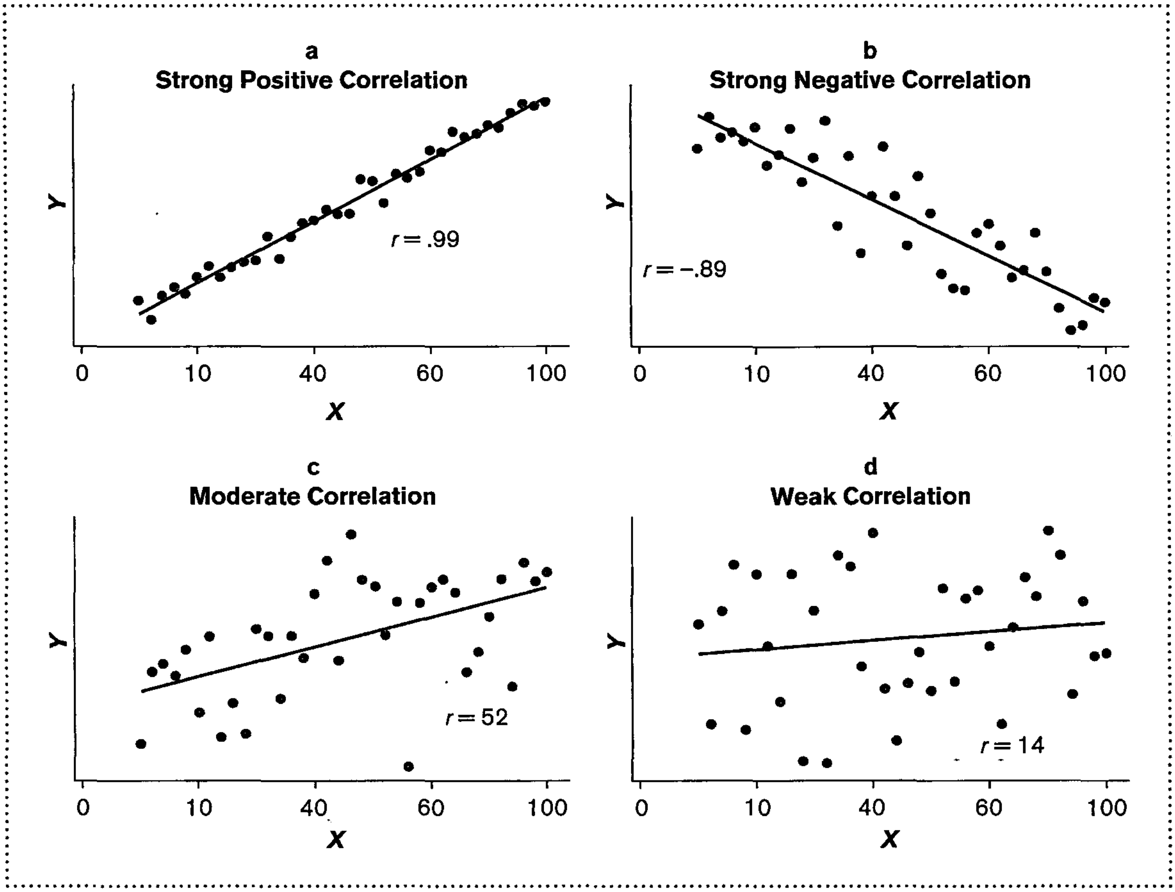
where Y_i and X_i are the values on Y and X of the ith observation; $\bar{Y}$ and $\bar{X}$ are the means of Y and X, respectively; $\hat{\sigma}_Y$ and $\hat{\sigma}_X$ are the sample standard deviations of Y and X, respectively; N is the sample size; and the summation is over all pairs of data points.

where the $\hat{\sigma}$ s are the sample standard deviations of X and Y. (As an aside, notice that r is partly a function of the size of the standard deviations. Given two samples with identical $\hat{\beta}$'s between Y and X, the one with the larger standard deviation, $\hat{\sigma}_X$, will *appear* to have the larger r and hence the larger linear correlation. But the magnitude of the relationship may simply be a function of the variability of X, not any intrinsic strength of the relationship.) We discuss the pros and cons of this feature of r in the next section.

Looking at the analysis of inequality, we see that the correlation between unionization and values of the Gini coefficients is –.63. This indicates a strong negative (linear) correlation.

To grasp the meaning of the numerical values of the correlation, try studying the patterns in figure 13-17. In graphs (a) and (b), most of the data lie near a straight line, and r is close to either its maximum value of 1 or its minimum value of –1. By contrast, graphs (c) and (d) show what you are likely to encounter in data analysis, moderate to weak relationships. Observe that the sign of r reflects the direction of the correlation. In each of these graphs, the data "behave very well" (this is a term statisticians commonly use): either there is a correlation or there isn't. Look, however, at figure 13-18. It illustrates a highly curvilinear relationship. A strong connection exists between Y and X—knowing the value of one would help you accurately predict values of the other—but the correlation coefficient (r = –.15) might suggest a weak relationship, until you remember that it measures the fit to a line. Once again, we stress the importance of examining scatterplots along with numbers. You are not likely to encounter such a strongly curved relationship in typical social science data, but it is important to remember that at this level of study, regression performs "best" when associations are linear or can be transformed to linearity.

FIGURE 13-17 Degree and Direction of Correlation

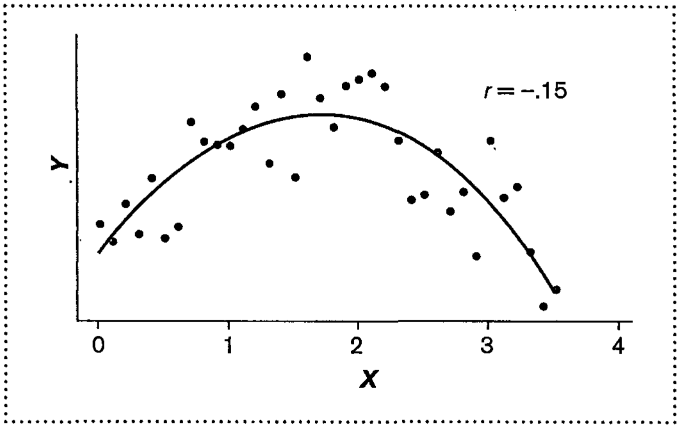


Standardized Regression Coefficients

The regression coefficient indicates how much Y changes—in values of the dependent variable—for a one-unit change in *X*. If you were relating years of education (*X*) to annual wages (*Y*), the regression parameter would be expressed in, say, dollars. As an example, an estimated coefficient of 1.25 should be thought of and interpreted as "There is a 1.25-unit increase in Y for every 1-unit increase in the independent variable." Earlier we found that the regression coefficient of Gini scores on union density equaled –0.14, or a 1 percent growth in union membership is associated with a decline (note the minus sign) of .14 units on the Gini scale. In many, if not most cases, this measure of association is exactly what is needed.

Yet social scientists sometimes prefer a "scale-free" statistic. For example, if independent variables had a common scale, their effect or impact could be compared unequivocally. (This statement rests on a major caveat discussed below.) Let's say a research team wants to explain variation in inequality. The problem with the regression coefficient is that a "one-unit change in X" means different things depending on the measurement scale. If X is income measured in dollars (a unit is \$1), the coefficient may have a very small numerical value; if it is measured in thousands of dollars (a unit is \$1,000), the coefficient will be larger.

FIGURE 13-18 **Curvilinear Relationship**



To solve this problem, researchers often rescale all variables so that a one-unit change has common meaning. The results are called **standardized variables**. To obtain them, you subtract the mean (of the variable) from each value and divide by the standard deviation. Consider a variable, X. Its sample mean is $\bar{X}$ and its standard deviation is $\hat{\sigma}_X$. Denoting the corresponding standardized value of X with lowercase x, the standardization is

$$x_i = \frac{(X_i - \bar{X})}{\hat{\sigma}_x}.$$

Keep a close eye on the symbols: lowercase letters denote standardized variables, whereas uppercase letters represent the raw data. After standardization, the unit of measurement becomes "standard deviations." Thus, a one-unit change in x is one standard deviation change. To clarify, think of union density as X; a one-unit change is a 1 percent change. Then standardize the variable to get x, standardized union density. A one-unit increase now means one standard deviation change.

The minimum union density in table 13-30 is 8.2 percent. When standardized, this becomes –1.29. Note the minus sign. The interpretation is that the minimum value lies 1.29 standard deviations *below* the mean. Similarly, the maximum is 2.09 standard deviations *above* the mean.

Standardized variables have several interesting and (for statisticians) important features. First, the mean and standard deviation of the recalibrated variables are always zero and 1.0, respectively. (Standardization is based on the fact that deviations from a mean add to zero.) Table 13-31 illustrates the technique for standardizing a variable and its properties.

TABLE 13-30 **Raw and Standardized Union Density Scores**

Example Values	Union Density	Standardized Union Density
Minimum	8.2%	–1.29
Mean	34.79%	0
Standard deviation	20.63%	1.0
Maximum	78%	2.09

More important, the advantage of standardization lies in the seeming simplicity of regression results. For example, when both *X* and *Y* are standardized and *y* is regressed on *x*, the resulting equation simplifies to

$$\hat{y}_i = \hat{\beta}^* x_i \,,$$

where $\hat{\beta}^*$ is the regression coefficient for the standardized data.

Notice that there is no constant (α): whenever *Y* and *X* have been standardized, the regression constant drops out. Also pay attention to $\hat{\beta}^*$. Called the **standardized regression coefficient**, this number is interpreted as in the usual way, except that now a one-unit change in *x* is a *one-standard-deviation change* in x. In other words, the independent variable's effect is measured in standard deviations of *y*, not the scale of the original dependent variable. A standardized coefficient has the same sign as its unstandardized cousin. The only difference is in numerical values and interpretations.

Incidentally, in two-variable regression, the standardized slope equals the correlation coefficient. Thus, it is not surprising that after standardizing the union and Gini variables, we find that $\hat{\beta}^*$ is –.63, the same value reported in the section on correlation.

The standardized regression coefficient tells us that a one-standard-deviation increase in union density corresponds to a .63-standard-deviation *decrease* in economic inequality. At the end of the day, though, the coefficients $\hat{\beta}$ and $\hat{\beta}^*$ differ numerically, but the overall picture they convey stays the same: the two variables are negatively correlated.

So, aside from simplifying the equation, why bother with standardized variables? Some social scientists believe that standardized regression coefficients enable you to rank the "relative importance" of independent variables on a dependent variable.[34] Imagine that you are trying to explain political participation. Your study includes (1) education in years of schooling, (2) annual family income in dollars, and (3) degree of partisanship measured on a 5-point scale (1 for "least partisan" to 5 for "highly partisan"). The literature tells you to expect a positive relationship between all three variables and the indicator of participation. But some authors say socioeconomic factors are more important explanations of political behavior than are political leanings; others disagree completely. So you perform an analysis and find the regression coefficients for education, income, and partisanship

[34] That is, $P + (1 - \text{P}) = 1$.

TABLE 13-31 Example of Raw Scores Converted to Standardized Values

Observation I	Raw Score (X_i)	Raw Deviation ($X_i - \bar{X}$)	Squared Deviation $(X_i - \bar{X})^2$	Standardized Score (x_i)	Squared Deviations $(x_i - \bar{x})^2 = x_i^2$
1	10	10 – 15 = –25	25	–1.34	1.79
2	12	12 – 15 = –3	9	–0.80	.64
3	14	14 – 15 = –1	1	–.27	.07
4	16	16 – 15 = 1	1	.27	.07
5	18	18 – 15 = 3	9	0.80	.64
6	20	20 – 15 = 5	25	1.34	1.79
Sums	90	0	70	0	5

Raw scores:

Mean of X: $\bar{X} = 90/6 = 15$.

Standard deviation of X: $\hat{\sigma}_X = \sqrt{70/5} = 3.74$.

Standardized scores:

Mean of x: $\bar{X} = 0/6 = 0$.

Standard deviation of x: $\hat{\sigma} = \sqrt{\frac{5}{5}} = 1.0$.

to be, respectively, .0001, .5, and 1. Numerically, $\hat{\beta}_{\text{partisanship}}$ is larger than either of the other two; hence, psychology seems more important than economic class. The problem is that since the variables have different measurement scales, the coefficients cannot be compared directly. If, however, you standardize all of the variables, the standardized coefficients might turn out to be .8, .5, and .2, in which case the socioeconomic variables seem to have the strongest relationship.

Although calculating standardized variables may be a good idea, their use in the preceding example works only if the independent variables are independent of one another, a situation that rarely arises in observational studies. In addition, even if you wish to take advantage of the standardized version, you should calculate the nonstandardized coefficient as well.

Inference for Regression Parameters: Tests and Confidence Intervals

This section builds on the ideas presented in chapter 12 and in previous sections regarding hypothesis testing. You may wish to review those topics briefly before proceeding.

Like any other statistical procedure, regression can be applied to sample or population data. In the present context, we assume that in a specified population a relationship exists between X and Y and that one way of describing it is with the regression coefficient β. This unknown quantity must be estimated with β, the sample regression coefficient. Briefly, we want to test the statistical hypothesis, H_0: $\beta_1 = 0$ (or some specified value) against H_A: $\beta_1 \neq 0$ (or, perhaps, $\beta_1 < 0$ or $\beta_1 > 0$). The test for its significance can go in two directions, both of which end up in essentially the same place. We describe the first here and save the other for the next chapter.

Under the assumptions stated at the beginning of the regression section (independent sampling, normally distributed errors, and so forth), we can test the null hypothesis that the constant and regression coefficient equal a particular value, typically zero. The estimated regression coefficient has a t distribution with $N - 2$ *df.* (When N becomes large—roughly 30 or more cases—the t blends into the standard normal distribution, for which a z statistic is appropriate.) The null hypothesis is usually simply H_0: $\beta_1 = 0$. This possibility is tested against a two-sided ($\beta \neq 0$) or one-sided ($\beta < 0$ or $\beta > 0$) alternative. The test statistic has the typical form—the estimated coefficient divided by the estimated standard error:

$$t_{\text{obs}} \frac{\hat{\beta}}{\hat{\sigma}_{\hat{\beta}}},$$

where $\hat{\sigma}_{\hat{\beta}}$ is the estimated standard error of the regression coefficient. Remember that if an estimator is calculated from many, many independent samples, these sample statistics will have a distribution, called the sampling distribution, with its own mean and standard deviation. When $\hat{\beta}$ is the statistic, its sampling distribution is the t distribution (the standard normal for a large N), which has mean β and standard error or deviation $\hat{\sigma}_{\hat{\beta}}$. Confidence intervals can be constructed in the usual way:

$$\beta \pm t_{(1-\alpha)/2.N-2\hat{\sigma}_{\hat{\beta}}},$$

where $t_{(1-\alpha)/2.\,N-2}$ is the value in appendix B that cuts off the upper $\alpha/2$ proportion of the distribution.

For example, earlier we estimated the regression of Gini scores on union density. The estimate turned out to be –.14, with a standard error of $\hat{\sigma}_{\hat{\beta}1} = .04$. It appears to be a small number, but is it statistically significantly different from zero? The sample size, $N = 21$, is less than 30, so we use a t distribution with $N - 2 = 21 - 2 = 19$ degrees of freedom to find the critical t at the .01 level, two-tailed test. It is 2.861. The observed t is

$$t_{\text{obs}} = \frac{\hat{\beta}_1}{\hat{\sigma}_1} = \frac{-.14}{.04} = -3.5.$$

This is a bit larger than the critical value, so we reject the null hypothesis at the .01 level and conclude that the population β_1 is probably not zero. Our best estimate is that it is about –.14. A test of the regression constant, β_0 is conducted in exactly the same manner: obtain the observed *t* statistic by dividing the estimated constant by its standard error. Recall that the regression constant for the inequality data is 36.44 with a standard error of 1.56, so the observed *t* is 36.43/1.56 = 23.35, which also greatly exceeds the critical value. Normally, regression results are displayed as in table 13-32. Symbols such as asterisks (*) are frequently used to denote the achieved significance, as shown in the table. (Sometimes they appear next to the names of the coefficients and sometimes next to the coefficients themselves, as in this table.)

Confidence intervals also appear in the table (in parentheses). They are found by

$$\begin{aligned}&\hat{\beta}_j \pm t_{\alpha/2=.01/2=.005,19}\hat{\sigma}_{\hat{\sigma}_j}\\&=-.14\pm 2.861(.04)\\&=-.25,-.03.\end{aligned}$$

Estimates of the constant usually do not have much practical meaning, but these days reporting them for quantities of interest (e.g., the regression coefficient that measures the impact of unionization on inequality) is required. Note that neither set of intervals includes zero, a fact consistent with the hypothesis test. As we explained in the last chapter, confidence intervals provide both a test of a statistical hypothesis and a range of plausible values for the coefficients.

TABLE 13-32 Regression of Gini on Union Density

Coefficient	Estimate (confidence intervals)	Standard Error	Observed *t*	Probability
Constant ($\hat{\beta}_0$)	36.44*** (31.98, 40.90)	1.56	23.41	.000
Coefficient ($\hat{\beta}_1$) Gini on union density	–.14** (–0.25, –0.03)	.04	–3.5	.002

Critical *t* for .01 level (two-tailed) with 19 *df*: 2.861.

** = significant at .01; *** = significant at .001.

Case Studies in Two-Variable Regression

We bring the chapter to a close by analyzing three additional examples. This analysis presents no really new ideas, but it does underscore our central theme: statistics requires more than the mechanical application of software to a pile of numbers. Instead, it requires clear thinking about what the data mean for the substantive problem. Doing so in turn requires a systematic approach:

- Examine each variable's summary statistics.
- Use graphs wherever possible and helpful.
- Always keep the units of analysis and measurement scales in mind.

CASE 1: LITERACY AND ECONOMIC DEVELOPMENT, A NONLINEAR RELATIONSHIP. The variables are GDP per capita (X) (measured in dollars)—an operational indicator of development—and literacy rates (Y) (percentage of adult population that is literate) in ninety-seven countries ranging from Angola to Zambia for the year 2004.[35] Table 13-33 shows the summary statistics.

Note these points:

- The difference between the median and mean GDP is quite large, with the mean being about two and a half times bigger. Also, three-quarters of the countries have GDPs below about $3,312 (see Q3), whereas the maximum is $34,340—a huge disparity. All of this adds up to the fact that the distribution is heavily skewed to the right.
- The opposite is true of literacy: it is skewed to the left. The minimum is just 19 percent, but in the bulk of the countries, at least two-thirds of the citizens can read. If fact, in one-quarter of the cases, literacy is virtually universal (i.e., third quartile = 96%). In addition, compare the median and mean: this time the mean is somewhat smaller, suggesting that a few low values are pulling the average down. The "typical" literacy rate is not 79 percent, as the mean suggests, but closer to 87 percent (see the median in the table). The conclusion? A relatively few nations have high levels of *il*literacy; most don't.
- GDP is measured in dollars. Therefore, a one-unit change in this variable doesn't amount to a hill of beans. To get a meaningful idea of the impact of GDP on education, we will be better off asking what a $500 or even $1,000 increase does.

[35] These sorts of data are widely available. For various (and irrelevant) reasons, we used Steven Finkel, Andrew Green, Aníbal Pérez-Liñán, Mitchell Seligson, and C. Neal Tate, *Cross-National Research on USAID's Democracy and Governance Programs—Codebook (Phase II)*. Available at http://www.pitt.edu/~politics/democracy/democracy.html

TABLE 13-33 **Summary Statistics for GDP and Literacy in Ninety-Seven Countries**

Variable	Minimum	Q1	Median	Mean	Q3	Maximum	Standard Deviation
GDP (in dollars)	$87.5	$67.41	$1,323	$3,181	$3,312	$34,340	$5,435.54
Literacy (percentage)	19.0%	67.4%	86.7%	78.6%	96.3%	99.8%	21.0%

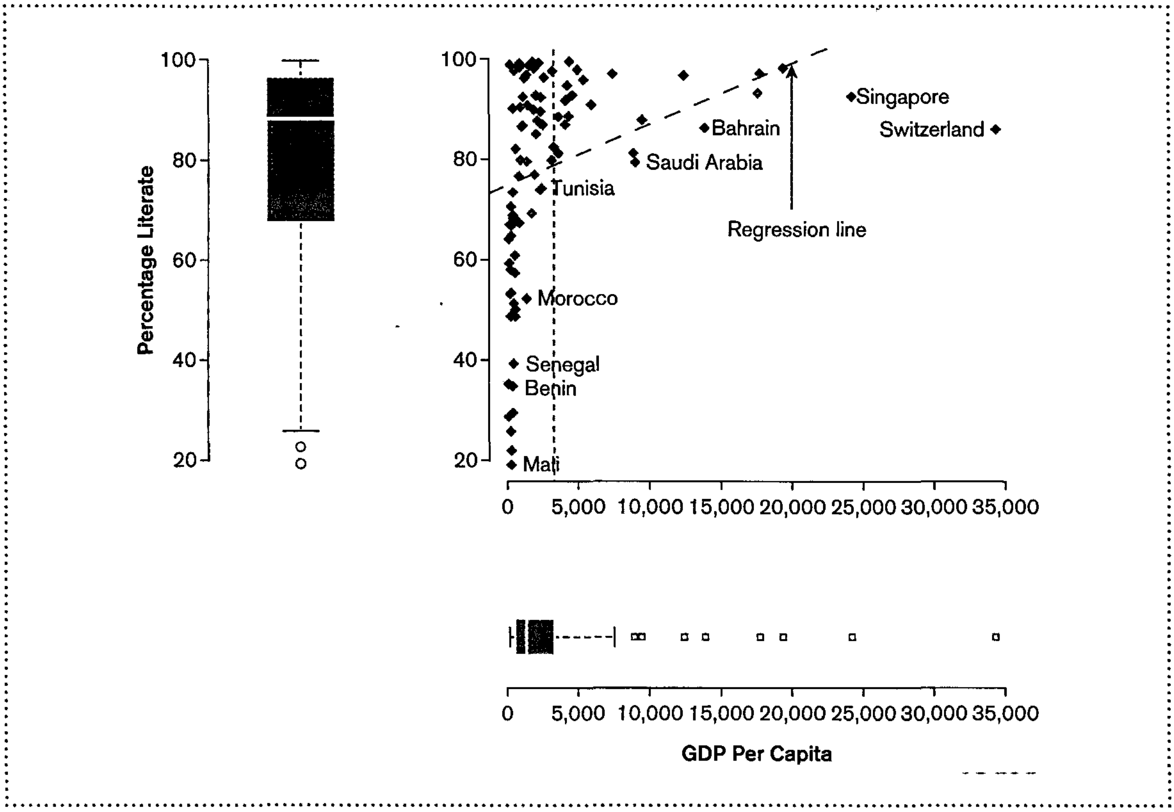
Source: Data from Steven Finkel, Andrew Green, Aníbal Pérez-Liñán, Mitchell Seligson, and C. Neal Tate, *Cross-National Research on USAID's Democracy and Governance Programs—Codebook (Phase II).* Available at http://www.pitt.edu/~politics/democracy/democracy.html

Now look at figure 13-19. The graph itself shows what at first sight looks like a very peculiar scatterplot. Some of the points have been identified to illustrate once again the basic idea of such a graph. It shows each unit's values on the two variables, literacy and per capita GDP. The skewness of the data can be seen in the accompanying boxplots. Remember, a boxplot gives a snapshot (but an informative one) of the data's main features. We see, for instance, that the median GDP is about $1,320. (The bar in the middle of the box stands for the median; read over to the scale to get its approximate value.)

The graph tells two stories. Or rather there are two substantive conclusions, one for the poorer nations and another for the richer ones. Have a look the plot of the points on the left of the dotted line, which marks off the first quartile. They seem stacked one atop the other. Let's try to tease some meaning out of the stack. In these countries, it appears that even relatively small increases in per capita GDP bring noticeably higher levels of literacy *or* possibly nations can boost reading skills while remaining quite poor. But after a certain point, no amount of income can raise literacy because the maximum is 100 percent. For example, even though Cyprus has a GDP per capita approximately one-third of Switzerland's, its literacy rate is about 10 points higher. This finding suggests that education may be affected by more than just material resources; surely culture, tradition, history, and social organization play a role. But the effect depends on the level of GDP. Such a situation is called "interaction," as we see in the following chapter.

What can we make of this relationship? Table 13-34 presents the results of the regression of literacy on income. The table is organized in a fashion common in scholarly publications. Below the estimated coefficients (in parentheses) are their standard errors. You can figure out the observed test statistic merely by dividing the standard error into the coefficient (e.g., 74.71/2.36 = t_{obs} = 31.66). Since there are 97 observations in the study, which is considerably more than 30, we can use the standard normal (z) distribution to find the probability that if the null hypothesis is true (H_0: β_0 = 0), we would get a z value of 31.66 or greater. It's practically zero.

FIGURE 13-19 **Literacy by GDP per Capita**



Source: Cross-National Research on USAID's Democracy and Governance Programs.

Hence, the highly significant test result. One conclusion is that overall there is not a linear but a curved or nonlinear correlation between the variables. Income's effects, in other words, are not constant across the full range of GDP. As development progresses, it has less and less effect on literacy. That's the substantive inference. A regression of literacy on GDP produces an estimated model: $\bar{Y}_i = 74.71 + .001GDP$.

The regression coefficient $\hat{\beta}_{\text{Literacy-GDP}} = .001$ looks tiny, but remember the measurement issue: GDP is measured in dollars (not in hundreds or thousands of dollars), so a one dollar increase or decrease is associated with less than 1 percent change in literacy levels. However, if we changed income by, for instance, $1,000, the effect on literacy would be 1,000 × .001 = 1 percent. At this point this still doesn't seem like such a big deal, but wait.

We also see that although both coefficients are highly significant (we do not accept the hypothesis that they are zero), the fit of the data to the model is weak—as measured by R^2, anyway. This lack of fit stems from the curved relationship.

TABLE 13-34 **Regression Results: Literacy on Income**

Variable	Raw Data	Logged Data
Constant	74.71*** (2.36)	9.56 (9.11)
GDP per capita	0.001** (0.00038)	9.66*** (1.25)

$R^2_{\text{raw}} = .10$ $R^2_{\text{logged}} = .39$.

** = significant at .01; *** = significant at .001

Remember that a basic assumption of regression analysis is that the relationship be linear. If it is not, a common response is to transform the data in such a way as to "straighten" them out by transforming the raw scores into a variable that will have a straight-line relationship with the dependent variable. Lots of rescalings are possible, and many tools exist for finding the optimal ones. Here we confine ourselves to taking the logarithm of GDP.[36] By the way, it should be mentioned that the logarithmic transformation of income data is quite common (almost mandatory) in economics and policy sciences.

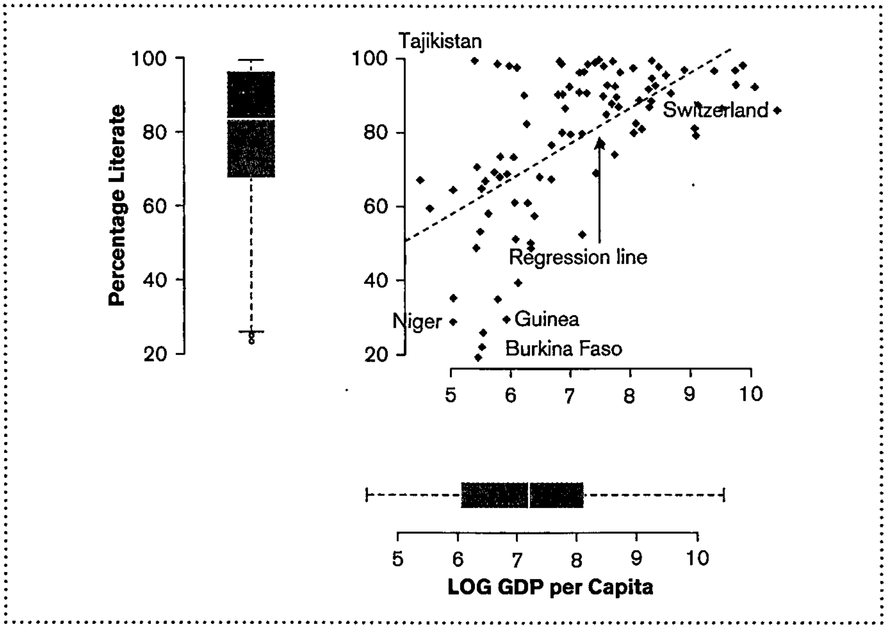
You can see the result of the transformation in figure 13-20. Notice first that literacy has been left untouched. But by using the *log* of GDP per capita, we have made its distribution much less skewed. (Compare the new boxplot with the previous one. The median and mean are now, respectively, 7.19 and 7.15 on the log scale.) More important, the relationship appears more linear. Notice, however, that at the lower end of the log scale, several countries lie a considerable distance from the estimated regression line. These points don't fit as neatly as we might expect, and perhaps warrant further investigation. The estimated coefficients, standard errors, and significance levels appear in the rightmost column of table 13-34. In this case, the coefficient for logged GDP is significant at the .001 level, but the constant no longer is. This once again forces us to ask if there is a substantive meaning for this term.

More generally, how does one know which variables should be altered, and by what method? There are systematic procedures for finding the best (most efficient) transformations. We cannot go into those techniques here. Instead, we suggest you engage in trial and error. Any measure of variable involving income will be a candidate for a logarithmic transformation. You can also try taking the square root of a variable or, going in the other direction, raise the variable to a power such as 1.5 or 2 (that is, for instance, $Y^{1.5}$ or Y^2).[37]

[36] We are using the natural logarithm.

[37] A first-rate and accessible (for undergraduate students in the social sciences) introduction to graphs, transformations, and data analysis is (if you can find it) Paul Velleman and David Hoaglin, *The ABC's of EDA: Applications, Basics, and Computing of Exploratory Data Analysis* (Duxbury, Mass.: Duxbury Press, 1981). Another good starting point is Frederick Hartwig and Brian E. Dearing, *Exploratory Data Analysis*, University Paper Series on Quantitative Applications in the Social Sciences, series 16 (Beverly Hills, Calif.: Sage, 1979).

FIGURE 13-20 **Literacy by Log GDP per Capita**



Source: Cross-National Research on USAID's Democracy and Governance Programs.

There are many other ways to tackle data like these. An appealing method is to treat the nations as a pool of developed and less developed economies. We could, for example, sort the countries into two groups based on whether they are below, at, or above the third income quartile. That is; those with GDPs greater than or equal to \$3,312 can be classified "Higher" and the others "Lower." By splitting the data this way, we can run two regressions, one for each group. The results, shown in table 13-35, confirm what the previous analysis pointed to as the interaction between wealth and literacy.

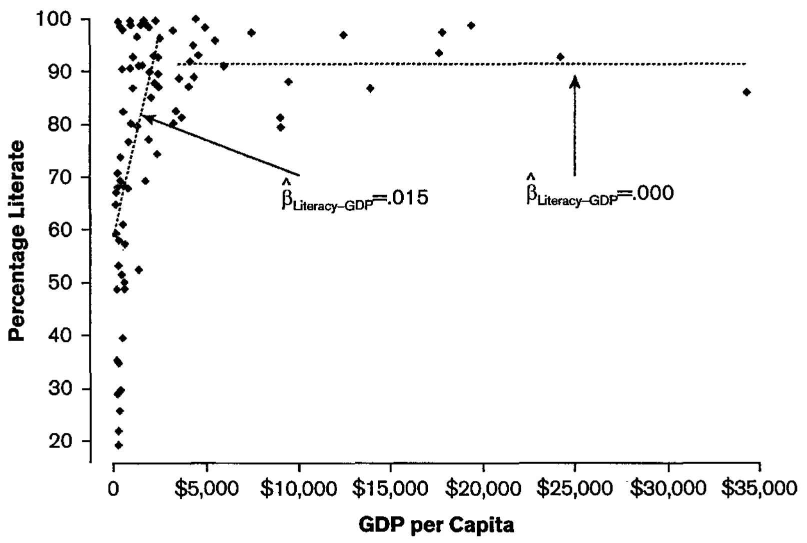
As should be apparent, each regression coefficient for the literacy data applied only to *lower* GDP countries ($N = 72$) is significant at the .001 level. Divide the estimates by their standard errors (in braces {}) to find the observed z statistics and use the z-table in appendix A to find the probability that if the betas were zero, you would obtain an observed z this large or larger. The confidence intervals tell you in an instant that neither set of limits contains zero and so the null hypotheses should probably be rejected. Figure 13-22 summarizes the previous analysis. We have, for example, printed the regression lines for both groups of countries. The one on the right is parallel to the x-axis because $\beta_{\text{Literacy-GDP}}$ is zero to three decimal places; a change in GDP does nothing for literacy in these countries.

TABLE 13-35 Regression Estimates for Two Groups

	Lower GDP ($N = 72$)	Higher GDP ($N = 25$)
Parameter	Estimate (99% confidence intervals) [standard error]	Estimate (99% confidence intervals) [standard error]
Constant, β_0	58.54*** (49.25–67.84) {3.51}	91.66*** (85.88–97.43) {2.05}
Regression, $\beta_{\text{Literacy-GDP}}$	0.015*** (0.008–0.023) {0.003}	.0000 (–0.0005–0.0004) {0.0002}

*** = significant at α = .001 level.

FIGURE 13-21 Two Regressions Fitted to Literacy Data

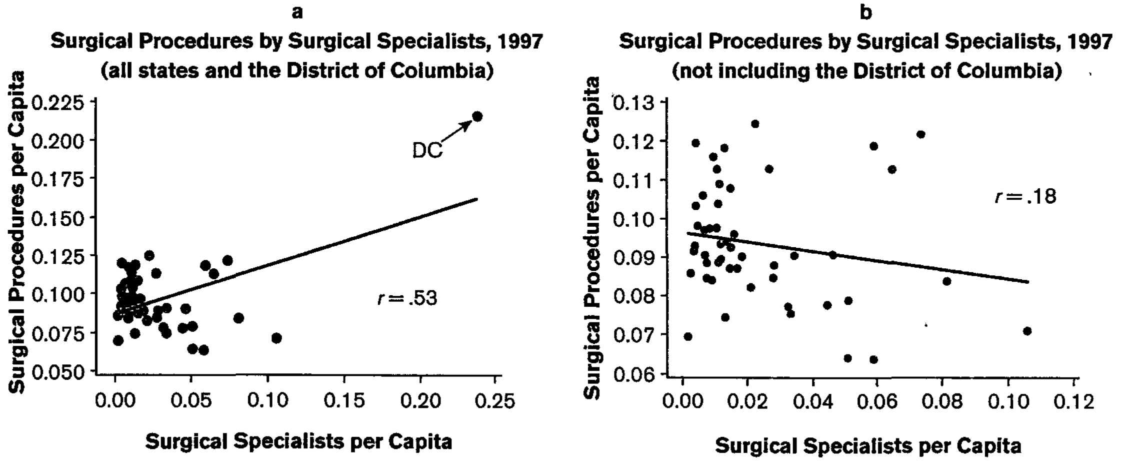


Source: Cross-National Research on USAID's Democracy and Governance Programs.

CASE 2: SURGICAL PROCEDURES AND PHYSICIAN AVAILABILITY—THE EFFECTS OF OUTLIERS. We now look at still another example of the importance of combing graphs, statistics, and general knowledge for obtaining substantively meaningful results. In order to do so, we look at an issue that plagues modern democracies, soaring health care costs. Just as the sample mean and standard deviation are sensitive to outlying or "deviant" values, so too are the regression parameter ($\hat{\beta}$) and correlation coefficient (r). The point is best demonstrated with an example.

The media sometimes report that health care costs continue to rise partly because Americans may be "overtreated." The argument runs as follows. To the extent that medical facilities and physicians are available, they will be utilized. Consequently, areas densely populated with, say, MRI devices will experience higher rates (per capita) of use than in places where they are scarce. Figure 13-22 takes a slightly different view of the problem. It shows the relationship between the number of surgical procedures carried out in the fifty states plus the District of Columbia by the number of surgical specialists. (Both variables have been converted to per capita indices.) The plot in figure 13-22a seems to back up the notion that more surgeons are accompanied by more surgeries. Naturally, it is possible that specialists migrate to places with a lot of sick people. On the other hand, it seems more probable that availability drives usage. The correlation coefficient (.53) supports a claim of a relatively strong positive correlation between operations and surgeons.

FIGURE 13-22 The Effect of an Outlier



Source: Health Care State Rankings, 1999 (Morgan Quitno Press).

Yet even a cursory glance at the distribution might raise suspicions, for one point lies far from all the others. It is the value for the District of Columbia, which has a high density of surgical specialists and operations. (The city is known for its many and famous health centers.) You may recall from chapter 11 that a "distant" point is called an outlier. Sitting far from all the other states, D.C. is a prime example of an outlier. Correlation and regression are functions of the sums squared deviations from means, and if one or two observations differ greatly from the others, their deviations will contribute a disproportionate amount to the totals. In technical terms, an outlier can exert great "leverage" on the numerical values of coefficients. That this is the case here can be seen in figure 13-22b. It displays the procedures by specialists for all the data *except* D.C. With that outlier removed—a valid, even necessary step in data analysis—the linear correlation disappears; the correlation coefficient $r = -.18$ has changed direction and moved into the "weak" range.

Statisticians know full well that standard regression models are not "robust" against problems such as leverage points and recommend paying attention to and adjusting for them. Luckily, most regression software offers the option of flagging these kinds of data.

CASE 3: JUDICIAL DECISION MAKING—INVESTIGATING LACK OF FIT AND RESIDUALS. We close the chapter by returning to judicial decision making. Remember the question of whether there is an association (possibly causal) between the nominating president's ideology and the ideological tenor of Supreme Court decisions. The analysis presented previously suggested why judicial nominations are the subject of bitter political disputes. Will we arrive at the same conclusion if we explore slightly different variables? Here we use the "Economic Liberalism Score of the Nominating President" as the independent variable and a "variable represents the percentage of 'liberal' votes cast by [justices] in the area of economics" as the dependent or response variable.[38] Both variables run from 0 to 100 percent, with larger numbers indicating greater liberalism. As before, we analyzed justices seated after 1950 and exclude two, John Roberts and Samuel Alito—both of whom had incomplete information on their voting records. The total *N* is 21.[39]

As usual, start with a scatterplot (figure 13-23). This is an ordinary plot that reveals a linear positive correlation between the two variables: the more "liberal" the nominating president, the more "liberal" the decisions. The slanted dotted line is the graph of the estimated regression. Note also that most points—14 out of 21, or about 71 percent—fall on the conservative side of the presidency scale. This, of course, demonstrates once again that more conservative and presumably

[38] Lee Epstein et al., "US Supreme Court Justices Database," 96, 114.

[39] The two latest appointees, Elena Kagan and Sonia Sotomayor, are not included for similar reasons.

Republican presidents have had a crack at filling vacancies. (Don't forget, though, that the analysis does not include the Bush or Obama administrations.) But if you study the vertical axis, you see that at least two (Warren and Brennan) of these more conservative appointees have (surprisingly?) liberal voting records, and one (Whittaker) is apparently far more conservative than the president who chose him. We have identified these individuals.

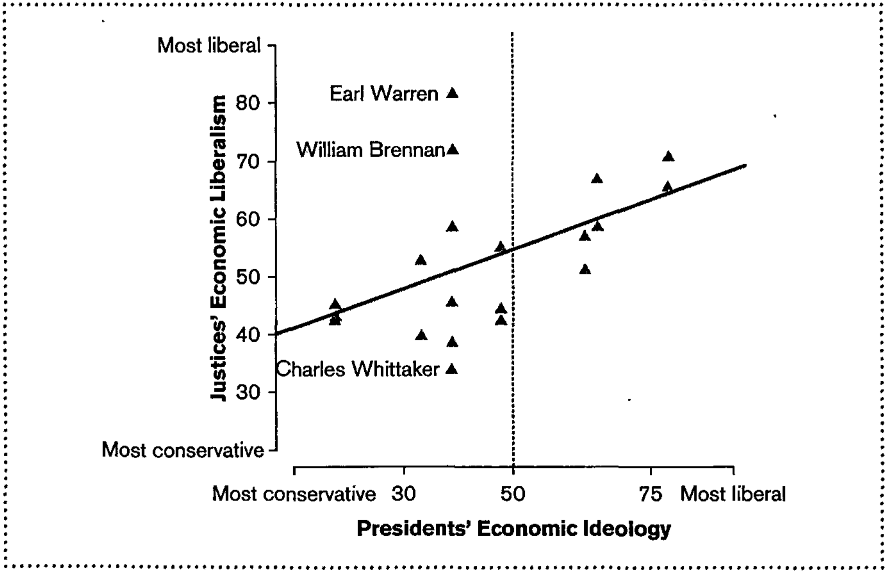
A statistical analysis lends support to the hypothesis of a positive linear connection: the simple correlation is $r = .51$, and the estimated regression coefficient is

$$\begin{array}{l} \hat{Y}_i = \underset{(6.32)}{37.69^{***}} + \underset{(.13)}{.34^{*}}\text{Presideology}, \end{array}$$

where the stars signify that both coefficients are statistically significant, one at the .001 level (the constant), the other at the .05 level (the regression coefficient).

The three labeled justices, however, might encourage us to take a closer look. First, we would recheck the data matrix to make certain the data have been correctly entered. More important, we might inquire into the circumstances of their nominations or into their backgrounds. Is there anything that might explain their anomalous positions? All three, for example, were nominated by the moderately conservative Dwight Eisenhower, a man without a strict ideological axe to grind.

FIGURE 13-23 **Does Presidential Ideology Affect Judicial Behavior?**



Source: Lee Epstein et al., "US Supreme Court Justices Database."

Thus, it is perhaps not too surprising that these justices' decisions span a wide spectrum. (Earl Warren had formerly been a Republican governor of California, and his subsequent rulings took some observers by surprise.) One might then wonder about the effects of these three people on the model's overall fit. (The R^2 is .26.) Would it help to treat these cases differently? Table 13-36 helps clarify the situation.

The table contains the raw data plus two important additional components: the predicted values of the dependent variable ($\hat{Y}_i$), found by applying the regression equation to the economic and presidential ideology scores, and the residuals, which measure the difference between observed and predicted Ys: $e_i = (Y_i - \hat{Y}_i)$. As explained under the discussion of how regression models are estimated, the residuals measure the line's "lack of fit." The table highlights the three largest residuals, which—no shocker here—belong to Warren, Brennan, and Whittaker.

What do you suppose would happen if these three observations were removed from the analysis? It should come as no surprise that all our measures of goodness of fit would improve (i.e., increase), as seen in table 13-37, especially the last column. Observe that even though their numerical values have changed hardly at all, all the terms in the estimated regression equation for the incomplete data are now highly significant. Usually when N is decreased, achieving significance becomes harder. Here, however, we have improved the fit by eliminating the three justices with the largest residuals.

Or have we? One might wonder about the propriety of selectively removing data that do not "agree" with one's hypothesis. Ordinarily there is little to recommend the practice. One of our goals here was to encourage a careful examination of the data in substantive terms to see why an estimated model may not fit what one thought was a good idea. Perhaps something about the Eisenhower presidency runs counter to the practice of basing judicial appointments on rigorous ideological tests. Our second objective is more pedagogical: we want to introduce the concept of residual analysis. Most statisticians use residuals to check assumptions and look for transformations that make these assumptions more tenable. But this is a vast and technical subject that we leave for the next chapter and further reading.[40]

Conclusion

This chapter has shown how to measure the existence, direction, strength, and statistical significance of relationships between two variables. We have emphasized the difference between association and causation. The particular techniques

[40] Most introductory texts on regression analysis devote chapters to model diagnostics and the exploration of residuals. One of our favorites is Thomas P. Ryan, *Modern Regression Methods* (New York: Wiley, 1997).

TABLE 13-36 Supreme Court Justices

Justice	President Ideology X_i	Economic Ideology Y_i	$\hat{Y}_i$	$e_i = (Y_i - \hat{Y}_i)$
Warren, Earl*	**38.8**	**81.59**	**50.37**	**31.22**
Harlan, John*	38.8	38.40	50.37	–11.97
Brennan, William*	**38.8**	**71.66**	**50.37**	**21.29**
Whittaker, Charles*	**38.8**	**33.78**	**50.37**	**–16.59**
Stewart, Potter*	38.8	45.52	50.37	–4.85
White, Byron	65.4	58.39	59.29	–0.90
Goldberg, Arthur	65.4	66.67	59.29	7.38
Fortas, Abe	78.2	70.67	63.58	7.09
Marshall, Thurgood	78.2	65.21	63.58	1.63
Burger, Warren	47.7	42.58	53.35	–10.77
Blackmun, Harry	47.7	55.02	53.35	1.67
Powell, Lewis	47.7	44.44	53.35	–8.91
Stevens, John	38.8	58.45	50.37	8.08
O'Connor, Sandra Day	17.6	43.17	43.26	–0.09
Rehnquist, William	17.6	45.05	43.26	1.79
Scalia, Antonin	17.6	42.31	43.26	–0.95
Kennedy, Anthony	17.6	44.66	43.26	1.40
Souter, David	33.1	52.55	48.46	4.09
Thomas, Clarence	33.1	39.71	48.46	–8.75
Ginsburg, Ruth Bader	63.1	56.78	58.52	–1.74
Breyer, Stephen	63.1	50.97	58.52	–7.55

*Nominated by Dwight Eisenhower.

Source: Lee Epstein et al., "US Supreme Court Justices Database."

used—cross-tabulation, difference-of-means test, analysis of variance, regression analysis, and correlation—all lead to inferential evidence but in no sense "prove" anything. This warning is especially apt for the examples presented above. Stripped as they are of the theoretical, contextual, and statistical rigor necessary to warrant being used as evidence, these examples have been presented merely as learning devices. What is especially important is to keep both technique and substance in proper perspective. One way of doing so is always to ask yourself, "What does this finding mean in the real world?" That we have tried to do by constantly going back and forth to the data and its real-world implications. We can, however, strengthen our case by adding additional variables.

TABLE 13-37 **The Effects of Case Deletion**

Dataset	Regression Equation	Measures of Fit
Full ($N = 21$)	$\hat{Y}_i = 37.69^{***} + .34^{*}$ Preseconomic.	$R^2 = .26$ $r = .51$
Incomplete ($N = 18$)	$\hat{Y}_i = 34.45^{***} + .37^{*}$ Preseconomic.	$R^2 = .59$ $r = .77$

* = significant at .05 level; *** = significant at .001 level.

TERMS INTRODUCED

Analysis of variance (ANOVA). A technique for measuring the relationship between one nominal- or ordinal-level variable and one interval- or ratio-level variable.

Chi square. A statistic used to test whether a relationship is statistically significant in a cross-classification table.

Correlation coefficient. In regression analysis, a measure of the strength and direction of the linear correlation between two quantitative variables; also called product-moment correlation, Pearson's *r*, or *r*.

Cross-tabulation. Also called a cross-classification or contingency table, this array displays the joint frequencies and relative frequencies of two categorical (nominal or ordinal) variables.

Degrees of freedom. A measure used in conjunction with chi square and other measures to determine if a relationship is statistically significant.

Difference-of-means test. A technique for measuring the relationship between one nominal- or ordinal-level variable and one interval- or ratio-level variable.

Direction of a relationship. An indication of which values of the dependent variable are associated with which values of the independent variable.

Effect size. How and how much a change in one variable affects another variable, often measured as the difference between one mean and another, often between a treatment group and control group.

Eta-squared. A measure of association used with the analysis of variance that indicates what proportion of the variance in the dependent variable is explained by variance in the independent variable.

Goodman and Kruskal's gamma. A measure of association between ordinal-level variables.

Goodman and Kruskal's lambda. A measure of association between one nominal- or ordinal-level variable and one nominal-level variable.

Kendall's tau-*b* and tau-*c*. Measures of association between ordinal-level variables.

Measure of association. Statistics that summarize the relationship between two variables.

Negative relationship. A relationship in which high values of one variable are associated with low values of another variable.

Null hypothesis. The hypothesis that there is no relationship between two variables in the target population.

Phi. An association measure that adjusts an observed chi-square statistic by the sample size.

Positive relationship. A relationship in which high values of one variable are associated with high values of another variable.

Proportional reduction in error (*PRE*) measure. A measure of association that indicates how much the knowledge of the value of the independent variable of a case improves prediction of the dependent variable compared to the prediction of the dependent variable based on no knowledge of the case's value on the independent variable. Examples are Goodman and Kruskal's lambda, Goodman and Kruskal's gamma, eta-squared, and R-squared.

Regression analysis. A technique for measuring the relationship between two interval- or ratio-level variables.

Regression coefficient. A measure that tells how much the dependent variable changes per unit change in the independent variable.

Residual. The numerical difference between an observed value and the corresponding value predicted by a model such as a regression equation.

***R*-squared.** The proportion of the total variation in a dependent variable explained by an independent variable.

Scatterplot. A graph that plots joint values of an independent variable along one axis (usually the *x*-axis) and a dependent variable along the other axis (usually the *y*-axis).

Somers' *D*. A measure of association between ordinal-level variables.

Standardized regression coefficient. A coefficient that measures the effects of an independent variable on a dependent variable in standard deviation units.

Standardized variable. A rescaled variable obtained by subtracting the mean from each value of the variable and dividing the result by the standard deviation.

Statistical independence. Property of two variables where the probability that an observation is in a particular category of one variable and a particular category of the other variable equals the simple or marginal probability of being in those categories.

Total variance. A numerical measure of the variation in a variable, determined by summing the squared deviation of each observation from the mean.

SUGGESTED READINGS

Achen, Christopher. *Interpreting and Using Regression.* Sage University Paper Series on Quantitative Applications in the Social Sciences, series no. 29. Beverly Hills, Calif.: Sage, 1982.

Agresti, Alan. *An Introduction to Categorical Data Analysis.* New York: Wiley, 1996.

Agresti, Alan, and Barbara Finlay. *Statistical Methods for the Social Sciences.* 3rd ed. Upper Saddle River, N.J.: Prentice Hall, 1997.

Faraway, Julian J. *Linear Models with R.* New York: Chapman & Hall/CRC, 2005.

Fox, John. *Applied Regression Analysis and Generalized Linear Models.* 2nd ed. Los Angeles: Sage, 2008.

Lewis-Beck, Michael S., ed. *Basic Statistics.* Vol. 1. Newbury Park, Calif.: Sage, 1993.

Velleman, Paul, and David Hoaglin. *The ABC's of EDA: Applications, Basics, and Computing of Exploratory Data Analysis.* Duxbury, Mass.: Duxbury Press, 1981.

CHAPTER 14

Multivariate Analysis

CHAPTER OBJECTIVES

14.1 Identify the tables or measures used to investigate the connection between variables.

14.2 Discuss how multivariate analysis is used to determine relationships between variables of categorical data.

14.3 Explain the tools that treat a dependent variable as a linear function of various combinations of independent variables.

14.4 Relate the ways in which to enter a categorical variable into a regression model.

14.5 Describe the process of logistic regression.

STUDIES CITED IN THE FIRST CHAPTER argue that one can identify two general causes for increased economic inequality in the United States, political and economic. Political changes include increase in business power at the expense of workers. Using their political clout, business elites, for example, prevailed on the political and economic system to limit the strength of labor unions. Other scholars disagree. Changes in inequality result from the international economic environment such as the growth of international trade and the demand for highly skilled as opposed to manual or routine workers. Neither business nor the political system can have much influence over these developments. Sociologists Bruce Western and Jake Rosenfeld summarize:

> During this time [1973–2007], wage inequality in the private sector increased by over 40 percent. Union decline forms part of an institutional account of rising inequality that is often

> contrasted with a market explanation. In the market explanation, technological change, immigration, and foreign trade increased demand for highly skilled workers. . . .[1]

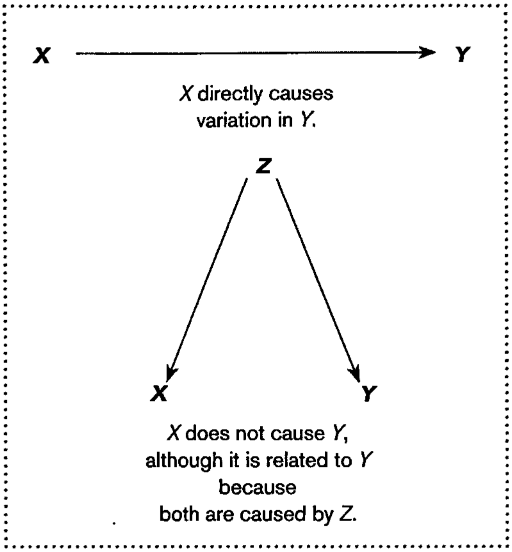
In the last chapter we described ways to measure the strength of association between two variables such as union strength and economic inequality. Usually, however, we want to know more than that: If there is there an association, why does it exist? Is it a causal linkage? Is it spurious? Is it part of a causal chain? Can our understanding of a dependent variable be expanded by adding other independent variables to the analysis? This takes us to "multivariate" analysis.

A first question is, Does adding a third variable to an analysis improve our understanding of an *X*-*Y* relationship? What happens if we add another independent variable? Do we know more than we did before, or is the extra information superfluous? This question is usually answered by comparing the fits of two models: one without the extra variable (the reduced model) and one with the variable added. Looked at this way, we say that one model is *nested* inside another. Suppose a model has just two variables, and *X* is used to predict *Y*. The variables are the set *XY*. Adding a third variable, *Z*, produces a larger set (*XYZ*), which contains the first one. The strategy is to contrast the adequacy of a model containing the full set of variables to one in which one or more variables have been eliminated—in other words, to compare the fit of the "full" model to the fit of the "reduced" model.

This method is also used to make causal inferences. If one can identify a connection between, say, *X* and *Y* that persists even after other variables (say, *W* and *Z*) have been taken into account, then there may be a basis for making a causal inference. We know by now that simply because a factor exhibits a strong relationship with a dependent variable, it does not follow that the former caused the latter. Both the independent and dependent variables might be caused by a third variable, which could create the appearance of a relationship between the first two. Only by eliminating this possibility can a researcher achieve some confidence that a relationship between an independent and a dependent variable is a causal one. Figure 14-1 illustrates the problem of distinguishing possible causal explanations.

[1] Bruce Western and Jake Rosenfeld, "Unions, Norms, and the Rise in U.S. Wage Inequality," *American Sociological Review* 76 (August 2011): 513.

FIGURE 14-1 **Causal and Noncausal Relationships**


X
Y
X directly causes variation in Y.
Z
X
Y
X does not cause Y, although it is related to Y because both are caused by Z.

As a start, we introduce the notion of "controlling" or holding a variable constant.

Holding a Variable Constant

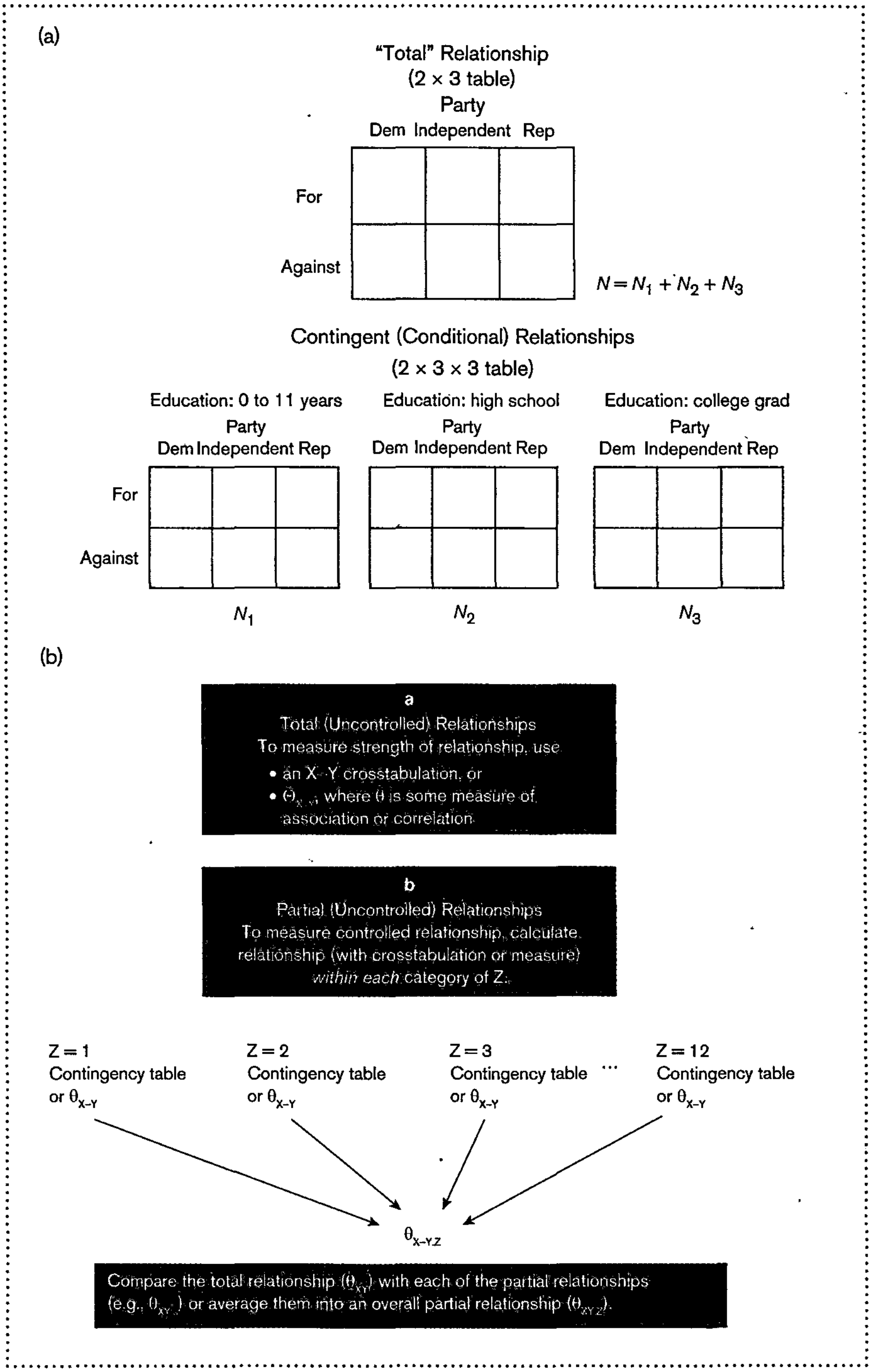
Suppose you are investigating the connection between two variables, *X* and *Y*. (*X* might be, say, party identification while *Y* is an attitudinal variable.) The first box in figure 14-2a represents a *total* or *original* association for a sample of size *N*. Depending on the variable scales, one can create a cross-tabulation and/or a single measure of association (θ) to investigate this total association. (See chapter 13 for different versions of θ.) Then, do the same thing within each category of another variable, *Z*, education perhaps. (See Figure 14-2b.) The set of tables or measures are the partial or controlled associations. In particular, they tell you how *X* and *Y* are related when *Z* equals a particular value. A summary of the partial tables will tell you how *X* and *Y* are related after *Z* has been controlled. We illustrate the idea with an example.

Multivariate Analysis of Categorical Data

Consider categorical data (nominal or ordinal scales). Suppose that we have hypothesized a relationship between attitudes toward government spending and presidential voting. Our hypothesis is that "the more a person favors a decrease in government spending, the more likely he or she is to vote Republican." Table 14-1 seems to confirm the hypothesis, since 64 percent of those who favored decreased spending voted Republican, whereas only 46 percent of those who favored keeping spending the same or increasing it voted Republican. This difference of 18 percentage points among a sample of 1,000 suggests that a relationship exists between attitudes toward government spending and candidate preferences.

At this point, you might ask, "Is there a causal relationship between opinion and vote (see the upper arrow diagram in figure 14-1), or does another factor, such as socioeconomic status (e.g., family income), create the apparent relationship?" Or, even if you are not interested in causality, the question arises: "Can the explanation of presidential voting be increased by including another variable?" After all, 36 percent of those who favored decreased spending (Democrats) voted contrary to the hypothesis, as did 46 percent of those in favor of maintaining or increasing spending levels (Republicans). Perhaps it would be possible to provide a better explanation for those voters' behavior.

FIGURE 14-2 **Total and Contingent Relationships**



(a)

"Total" Relationship
(2 × 3 table)

	Party Dem	 Independent	 Rep
For			
Against			

$N = N_1 + N_2 + N_3$

Contingent (Conditional) Relationships
(2 × 3 × 3 table)

Education: 0 to 11 years

	Party Dem	 Independent	 Rep
For			
Against			

N_1

Education: high school

	Party Dem	 Independent	 Rep
For			
Against			

N_2

Education: college grad

	Party Dem	 Independent	 Rep
For			
Against			

N_3

(b)

a
Total (Uncontrolled) Relationships
To measure strength of relationship, use
- an X–Y crosstabulation, or
- Θ_{X-Y}, where θ is some measure of association or correlation

b
Partial (Uncontrolled) Relationships
To measure controlled relationship, calculate relationship (with crosstabulation or measure) *within each* category of Z.

Z = 1
Contingency table
or θ_{X-Y}

Z = 2
Contingency table
or θ_{X-Y}

Z = 3
Contingency table
or θ_{X-Y}

…

Z = 12
Contingency table
or θ_{X-Y}

$\theta_{X-Y.Z}$

Compare the total relationship (θ_{XY}) with each of the partial relationships (e.g., $\theta_{XY.}$) or average them into an overall partial relationship ($\theta_{XY.Z}$).

TABLE 14-1 **Relationship between Attitudes toward Government Spending and Presidential Vote**

	Independent Variable: Attitudes toward Government Spending		
Dependent variable: presidential vote	**Decrease spending**	**Keep spending the same or increase it**	***(N)***
Republican	64%	46%	(555)
Democratic	36%	54%	(445)
Total	100%	100%	
(*N*)	(550)	(450)	(1,000)

Note: Hypothetical data.

A second independent variable that might affect presidential voting is personal income. People with higher earnings might favor decreased government spending because they feel they gain little from most government programs.[2] Those with higher incomes might also be more likely to vote Republican because they perceive the GOP as supporting decreases in government spending. By the same token, people having lower incomes might feel both that increased government spending would help them *and* that Democrats generally support their interests. Therefore, income might influence both attitudes toward government spending and presidential voting, thus creating the appearance of a relationship between the two.

To consider the effect of income, we need to bring it explicitly into the analysis and observe the resulting relationship between attitudes and voting. In a **multivariate cross-tabulation**, we **control** for a third variable **by grouping**; that is, we group the observations according to their values on the third variable and then observe the relationship between opinions on spending and voting within each of these groups. In our example, each group consists of people with more or less the same income. Therefore, if a relationship between opinions on spending and voting in these groups remains, it cannot be due to income.

Table 14-2 shows what might happen were we to control for income by grouping respondents into three income levels: high, medium, and low. Notice that it contains three contingency tables: one for each category of income, the control variable. Within each of the categories of income there is now *no* relationship between spending attitudes and presidential voting. Regardless of their attitudes on spending, 80 percent of respondents with high incomes voted Republican, 60 percent

[2] See Benjamin I. Page, Larry Bartels, and Jason Seawright, "Democracy and the Policy Preferences of Wealthy Americans," *Perspectives on Politics* 11 (March 2013): 51–73.

with medium incomes voted Republican, and 30 percent with low incomes voted Republican. Once the variation in income was removed by grouping those with similar incomes, the attitude-vote relationship disappeared. Consequently, income is a possible alternative explanation for the variation in presidential voting.

The original relationship, then, was spurious. Remember that a spurious relationship is one in which the association between two variables is caused by a third. Note, however, that these remarks do not mean that there is *no* relationship between spending attitudes and presidential voting, for there is such a relationship, as table 14-1 shows. But this original relationship occurred only because of the variables'

TABLE 14-2 Spurious Relationship between Attitudes and Presidential Voting When Income Is Controlled

Control Variable: Income	Independent Variable: Attitudes toward government spending		
Dependent Variable: presidential vote	**Decrease spending**	**Keep spending the same or increase it**	**(*N*)**
High income			
Republican	80%	80%	(240)
Democratic	20%	20%	(60)
Total	100%	100%	
(*N*)	(250)	(50)	(300)
Medium income			
Republican	60%	60%	(210)
Democratic	40%	40%	(140)
Total	100%	100%	
(*N*)	(200)	(150)	(350)
Low income			
Republican	30%	30%	(105)
Democratic	70%	70%	(245)
Total	100%	100%	
(*N*)	(100)	(250)	(350)

Note: Hypothetical data.

relationships with a third factor, income. Thus, spending attitudes cannot be a cause of presidential voting because within income groups, they make no difference whatever. (See the lower arrow diagram in figure 14-1.)

Because we have been using hypothetical data, we can easily illustrate other outcomes. Suppose, for instance, the control variable had absolutely no effect on the relationship between attitudes and vote. The result might look like the outcomes in table 14-3. We now see that the strength and direction of the relationship between attitudes and voting are the same at all levels of income. In this situation, members of the upper-income group behave just like those in the lower levels. Given these data, we might be tempted to support the argument that attitudes toward government spending are causally related to candidate choice. But, of course, a critic could always say, "But you didn't control for Z." That would be a valid statement, provided the skeptic provided a plausible reason why Z would have an effect on the original relationship. A randomized controlled experiment, in contrast to an observational study, theoretically eliminates all alternative explanatory variables at one fell swoop.

TABLE 14-3 Relationship between Attitudes and Presidential Voting after Income Is Controlled

Control Variable:			
Income	Independent Variable: attitudes toward government spending		
Dependent Variable: presidential vote	Decrease spending	Keep spending the same or increase it	(*N*)
High income			
Republican	64%	46%	(183)
Democratic	36%	54%	(117)
Total	100%	100%	
(*N*)	(250)	(50)	(300)
Medium income			
Republican	64%	46%	(197)
Democratic	36%	54%	(153)
Total	100%	100%	
(*N*)	(200)	(150)	(350)
Low income			
Republican	64%	46%	(179)
Democratic	36%	54%	(171)
Total	100%	100%	
(*N*)	(100)	(250)	(350)

Note: Hypothetical data.

These hypothetical data illustrate ideal situations. Consider, then, an actual multivariate cross-tabulation. Political pundits and campaign strategists, for example, are preoccupied with geographical variation in attitudes and voting. They talk of "blue" (Democratic) and "red" (Republican) states to describe typical voting patterns in these areas. Let's investigate regional differences regarding an ongoing "cultural" or social issue, prayer in public schools. To start, we created a "region" variable by combining respondents in the 2008 General Social Survey into four groups: (1) the "coasts," which include the Pacific, New England, and Mid-Atlantic states; (2) the "industrial" upper Midwestern states; (3) the traditional or Deep South; and (4) a conglomeration of south Atlantic and mountain states, which we label simply the "extended Sun Belt." The first two generally support Democrats for president and are thought to be centers of "liberalism." The remaining two are commonly identified with conservative and Republican voting patterns. (Needless to say, there is a lot of heterogeneity in these groupings; we use them merely for illustrative purposes.) Table 14-4 shows how people in different regions think about Supreme Court rulings limiting prayer in public schools. (For simplicity's sake, we have recoded the responses to "yes, favor" and "no, do not favor" prayer in the classroom.) The variation in the percentages saying "no" suggests an effect of region on public opinion. More than half of those on the coasts approve of the Court's decision, while only a quarter of those in the South do. The other regions fall in between. What, if anything, accounts for these differences?[3]

TABLE 14-4 **Total Relationship between Region and School Prayer**

Prayer in public schools okay?	Generally Democratic (Blue) States		Generally Republican (Red) States	
	East and West Coast	Industrial North Central	Extended "Sun" Belt	Traditional South
Yes	46.4%	59.8%	59.7%	75.8%
No	53.6%	40.2%	40.3%	24.2%
	100%	100%	100%	100%
	(386)	(275)	(423)	(199)

N = 1283; chi square = 47.9, 3 *df*; prob = .000; phi = 0.19.

Question: "The United States Supreme Court has ruled that no state or local government may require the reading of the Lord's Prayer or Bible verses in public schools. What are your views on this—do you approve or disapprove of the court ruling?"

Source: James A. Davis, Tom W. Smith, and Peter V. Marsden, *General Social Surveys*, 1972–2008, Roper Center for Public Opinion Research, University of Connecticut/Ann Arbor, Mich.: Inter-university Consortium for Political and Social Research.

[3] Totals do not add exactly across tables because (1) some observations have missing values on religiosity as well as opinion on prayer in public schools, and (2) weighted data were used in the analysis and small rounding errors occur.

More precisely, is there something about a geographical area that induces people to think one way or another? Or—more likely—do different regions contain different kinds of voters, and do these characteristics—not geography, per se—explain variation in opinions? Since the South stands out so much and we are dealing with a religious issue, an obvious candidate variable to add to the mix is some kind of indicator of religiosity. After all, the deep south was familiarly known as the "Bible Belt," and even today it is thought of as a stronghold of Christian conservatism. Therefore, let's include "fundamentalism" in the analysis. The GSS survey contains an item, "Fundamentalism/liberalism of respondent's religion," to which responses are coded "fundamentalist," "moderate," and "liberal"; the latter category presumably includes atheists, agnostics, and skeptics, as well as religious people who nevertheless do not take sacred texts literally. Table 14-5 shows a multiway table in which the original region-opinion relationship is examined for each of the three levels of fundamentalism.

To make sense of the data, we need to explore each subtable individually and carefully. Look first, then, at the fundamentalists (table 14-5a). The overwhelming majority of respondents in *each* region favor allowing prayer in schools. The percentages run from 67 to more than 80 percent. There are differences, to be sure—the fundamentalists on the "coasts" appear to be a bit more secular than their counterparts elsewhere. Nonetheless, the relationship is rather weak. The same is true for moderates (the middle table), although the proportions saying "no" are somewhat larger. Finally, we see that the region-attitude association is strongest and clearest in the last category, "liberals." Except in the South, a majority of respondents oppose organized prayer reading in public education. But opposition declines as one moves across the table.[4]

Further insight is achieved by looking at the chi-square statistics in each table and as compared to the overall chi square in table 14-5. They seem to indicate a weak to nil association in the first two levels of fundamentalism and a moderate one in the third table. We see, for instance, that even among nonfundamentalists in the South, there is solid backing for school prayer (67%) but not so on the coasts, where the opposition exceeds 70 percent. So our overall conclusions might be that (1) there are regional differences in attitudes, and (2) these differences are partly explained by one's degree of religious commitment.

Using summary statistics such as categorical measures of association or observed chi-square statistics helps because we can quickly average them across tables. The overall chi square for table 14-4 is 47.9 with 3 degrees of freedom; the weighted (by

[4] Notice that lambda in this table is zero. You may recall from the previous chapter that lambda will equal zero whenever the modal marginal category of *Y* is also the mode in each level of *X*.

TABLE 14-5 Controlled or Contingent or Conditional Relationships

a. Religiosity = fundamentalist				
Prayer in public schools	The "Coasts"	Industrial North	Sun Belt	Traditional South
No	32.9%	15.4%	25.2%	20.0%
Yes	67.1%	84.6%	74.8%	80.0%
	100%	100%	100%	100%
	(51)	(64)	(143)	(85)

N = 343; chi square = 5.68, 3 *df;* prob = .13; phi = 0.13.

b. Religiosity = moderate				
Prayer in public schools	The "Coasts"	Industrial North	Sun Belt	Traditional South
No	47.1%	39.9%	41.8%	26.0%
Yes	52.9%	60.1%	58.2%	74.0%
	100%	100%	100%	100%
	(181)	(117)	(125)	(75)

N = 498; chi square = 9.92, 3 *df;* prob = .02; phi = 0.15.

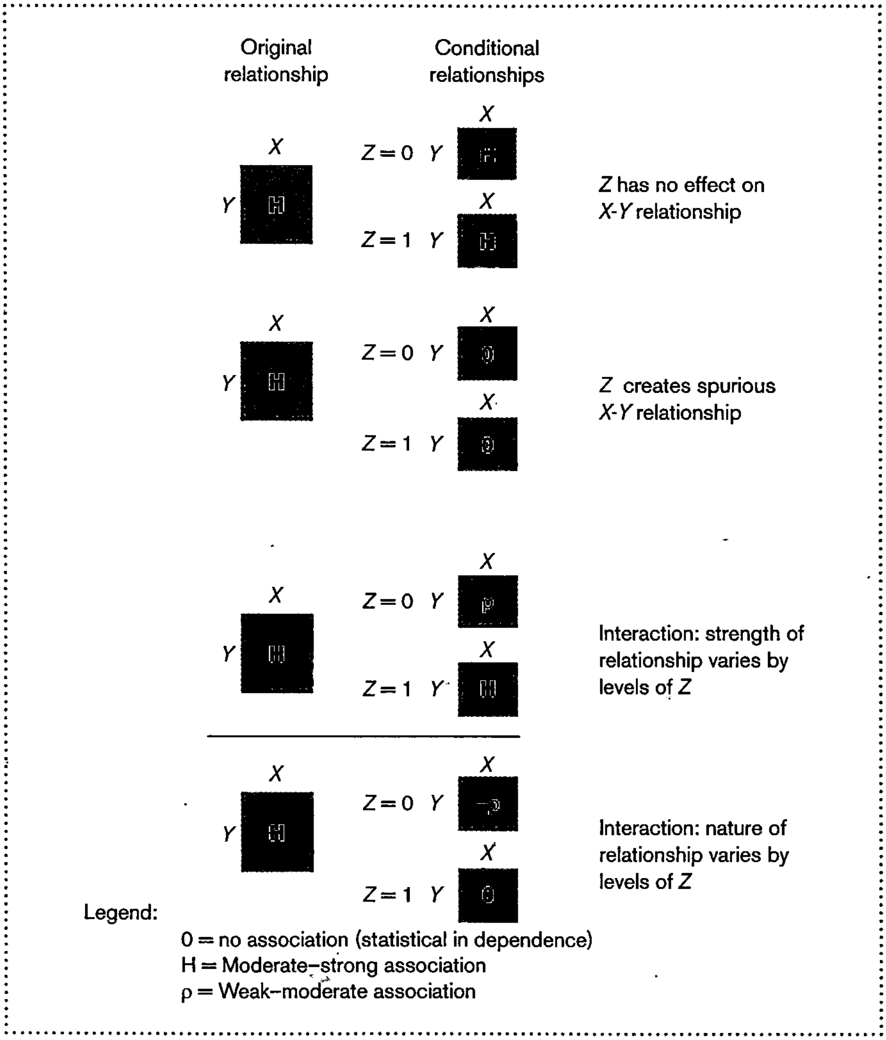
c. Religiosity = liberal				
Prayer in public schools	The "Coasts"	Industrial North	Sun Belt	Traditional South
No	73.0%	59.3%	52.2%	32.6%
Yes	27.0%	40.7%	47.8%	67.4%
	100%	100%	100%	100%
	(132)	(89)	(140)	(34)

N = 395; chi square = 23.06, 3 *df;* prob = .000; phi = 0.24.

Source: Data from table 14-4.

number of cases in each subtable, N_i) average of chi squares in table 14-5 is 12.2, again with 3 degrees of freedom. So the "controlled" relationship seems weaker than the total association. The average of the phi coefficients (remember that phi is the square root of the observed chi square divided by the sample size) is a tad smaller than the value in the main table (.17 versus .19).[5]

FIGURE 14-3 **How to Interpret Conditional Relationships**


Original relationship
Conditional relationships
X
Y
H
Z=0 Y
X
H
Z=1 Y
X
H
Z has no effect on X-Y relationship
X
Y
H
Z=0 Y
X
0
Z=1 Y
X
0
Z creates spurious X-Y relationship
X
Y
H
Z=0 Y
X
ρ
Z=1 Y
X
H
Interaction: strength of relationship varies by levels of Z
X
Y
H
Z=0 Y
X
−ρ
Z=1 Y
X
0
Interaction: nature of relationship varies by levels of Z
Legend:
0 = no association (statistical in dependence)
H = Moderate–strong association
ρ = Weak–moderate association

[5] There are techniques for "partitioning" a slightly different version of the chi square into components—one for each table—that add up to the total chi square.

Admittedly, this sort of analysis requires absorbing a lot of numbers and trying to discern patterns among them. Here are some guidelines, although in a moment we present a more formal procedure. (Figure 14-3 may help.)

Keep separate in your mind the original, uncontrolled relationship, *X*-*Y*. The goal is to see what happens to it when additional variables are introduced.

- If at each level or value of the conditioning variable, *Z*, there are approximately the same kind and degree of connection between *X* and *Y* as appear in the original, then *Z* may not be relevant to the *X*-*Y* association.
- Are the controlled relationships on average weaker or smaller than the original? If so, *Z* may be a (partial) spurious cause of the *X*-*Y* relationship, *or* there may be a spurious relationship or maybe a "causal sequence": that is, $X \rightarrow Z \rightarrow Y$. (Controlling for *Z* in either case reduces or eliminates the *X*-*Y* association.)
- Is the relationship between *X* and *Y* strong at some levels of *Z* but not others? If so, there may be statistical **interaction**. Interaction means that the strength, direction, and nature of the *X*-*Y* relationship depend on levels of the control variable. At the high end of the *Z* scale, there may be little or no connection between *X* and *Y*, while in the middle there is a negative correlation and there is a modest negative relationship for those cases with low values on *Z*. If interaction exists, the impact of *X* on *Y* depends on another variable and merits careful scrutiny. Such activity is sometimes referred to as "specifying" the relationship.

Social scientists have generally moved away from the analysis of multivariate cross-tabulations using percentages and measures of association. A variety of sophisticated and powerful techniques have been developed to describe complex contingency tables with "parsimonious" models.[6] There are two general approaches. Many sociologists, biometricians, demographers, and economists have developed methods designed explicitly to tease out of cross-tabulations as much information as possible. Another, more widely adopted method in political science and other fields is to apply generalized linear models that include combinations of quantitative and qualitative data.[7] We conclude this section by comparing the analysis of cross-tabulations with the randomized experimental design discussed in chapter 6. The goal of the latter is to see if one factor causes another. By randomly assigning individuals to treatment (experimental) and control groups, the investigator (in theory at least)

[6] An excellent introduction is Alan Agresti, *Categorical Data Analysis*, 2nd ed. (New York: Wiley, 2002). Another very useful book is Bayo Lawal, *Categorical Data Analysis with SAS and SPSS Applications* (Mahwah, N.J.: Erlbaum, 2003).

[7] The seminal work is Leo A. Goodman (with Clifford C. Clogg), *The Analysis of Cross-Classified Data Having Ordered Categories* (Cambridge, Mass.: Harvard University Press, 1984).

can scrutinize a relationship between X and Y uncontaminated by other variables, such as Z. In most research settings, however, randomization is simply not possible. Given a hypothesis about voter turnout and social class, for instance, how could a researcher randomly place someone in a particular occupation and then wait to see what effect this placement had on the person's behavior? Therefore, instead of using randomization to get rid of potentially contaminating variables, it is necessary to try to control for them manually. That is, the investigator has to explicitly identify variables (for example, Z) that might be influencing the X-Y relationship, measure them, and then statistically control for them just as we did in table 14-5. In that case, we looked at the association between the variables *within* levels of the third factor. This approach is possible if the control factor is categorical and the total number of cases is large. Other techniques are needed for different circumstances. In the next section, we discuss the cases of one continuous dependent variable and two or more categorical test factors.

Linear Models

The analysis and interpretation of multidimensional contingency tables like those just presented are complicated because so much information has to be gleaned from so many cell percentages and column totals. We now move to a framework and set of tools that overcome those and other problems. More important, this analytic structure extends to mixtures of categorical and numeric data. The result is that we treat a dependent variable as a (linear) function of various combinations of nominal-, ordinal-, and interval-scale independent variables.

Multiple Regression Analysis

When the dependent variable is measured at the interval or ratio level, we usually use **multiple regression analysis** to investigate how its values are affected by two or more independent variables. Chapter 13 noted that the aim of regression analysis is to propose and estimate a model (an equation), $Y_i = \beta_0 + \beta_1 X + \varepsilon_i$ that in some sense best describes or summarizes how X and Y are related. Finding the "best" model is called "fitting" the data, and predicted scores are called "fitted values." Recall in addition that a regression coefficient, which lies at the core of the model, tells how much the dependent variable, Y, changes for a one-unit change in the independent variable, X. Regression analysis also allows us to test various statistical hypotheses such as $\beta_1 = 0$, which means there is no linear relationship between an independent and a dependent variable. A regression equation, moreover, may be used to calculate the predicted value of Y for any given value of X. And the residuals or distances between the predicted and observed values of Y lead to a measure (R^2) of how well the equation fits the data.

As the name implies, multiple regression simply extends these procedures to include more than one independent variable.

The general form of a linear multiple regression equation is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_K X_K + \varepsilon.$$

Let's examine this equation to make sure its terms are understood. In general, it says that the values of a dependent variable, Y, are a *linear* function of the values of a set of independent variables.[8] The function is linear because the effects of the variables are additive. How the independent variables influence Y depends on the numerical values of the βs.

As in previous chapters, parameters are denoted by lowercase Greek letters. The first beta (β_0) is a **regression constant.**[9] It can be interpreted in many ways, the simplest being that β is the value of Y when all the independent variables have scores or values of zero. (Just substitute zero for each X and note that all the terms except the first drop out, leaving $Y_i = \beta_0 + \varepsilon_i$.)

Each β in the equation is called a **partial regression coefficient** because it indicates the relationship between a particular X and the Y *after all the other independent variables have been "partialed out" or simultaneously controlled.* The presence of ε (epsilon), which stands for error, means that Y is not a perfect function of the Xs. In other words, even if we knew the values of every X, we could not completely or fully predict Y; there will be errors. (In the symbols used in chapter 13, we denoted this idea by $Y - \hat{Y}$.) But regression proceeds on the assumption that the errors are random or cancel out and that their average value is zero. Hence, we can rewrite the regression equation as

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_K X_K.$$

Read this last equation as "The expected value of Y is a linear (or additive) function of the Xs."

Finally, predicted values of Y (denoted $\hat{Y}$) may be calculated by substituting any values of the various independent variables into the equation. With these predictions in hand, one can also compute residuals. It works this way:

- Predicted values: $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \ldots + \hat{\beta}_K X_{Ki}$
- Residuals: $\hat{\varepsilon}_i = (Y_i - \hat{Y}_i) = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \ldots + \hat{\beta}_K X_{Ki})$

[8] Sometimes the term *additive* is used to describe the models.

[9] In chapter 13, we also called this term the *intercept* because it has a simple geometric interpretation.

HELPFUL HINTS

Assumptions of Linear Models

For the estimation, testing, and hence interpretation to make sense, certain conditions have to be met, at least to one degree or another. The assumptions are extensions of those presented in chapter 13.

- Independent observations
- No measurement error in the Xs
- Correct model specification
  - All relevant *X*s are included. (Irrelevant independent variables will inflate the prediction error but not mess up the estimators.)
  - The dependent variable is a linear function of the independent variables.
  - A current value of *Y* is not a function of previous values of *Y*. Violations of this assumption are likely to be present in time series data.
- Multicolinearity
  - No *X* is an exact linear function of another *X* or *X*s. For instance, if $X_1 = 2X_2$, the variables are said to be *collinear*, and the estimation process breaks down. If X_i, X_j are highly but not perfectly correlated, the estimates of relevant betas may be suspect.
- Error term, ε
  - The expected value of the errors is 0: $E(\varepsilon_i) = 0$.
  - Constant variance (Homoscedasticity): The variation in *Y* is the same at each level of *X*.
  - Errors are uncorrelated with any *X*. That is, linear model analysis assumes that, say, $\rho X_{\text{rhoX}_i e_i} = 0$ where $\rho_{\text{rhoX}_i e_i}$ is the correlation of X_i and the error.
  - Errors are mutually independent: The correlation of ε_i with ε_j is zero.

Textbooks are loaded with advice and techniques to check the tenability of these assumptions. It is often possible to take corrective action when data do not seem to meet the requirements, but for the most part we will just take them for granted.

Check out more Helpful Hints at **edge.sagepub.com/johnson8e**

Residuals, which estimate the errors (ε_i) in the model, are an important tool in verifying assumptions and gauging the fit of the model to the data, as we see below.

Interpretation of Parameters

So important are the partial regression coefficients that we should examine their meaning carefully in the context of a couple of specific examples. First, let's briefly return to the problem of inequality. In the previous chapter, we discovered that union membership is negatively correlated with income inequality as measured by the Gini coefficient. Suppose we want to extend the analysis by adding an additional explanatory variable. Since we are interested in how workers are able to influence the distribution of wealth through the mobilization of labor power, a sensible candidate for inclusion is an indicator of "employment protection," which the Organisation for Economic Co-operation and Development (OECD) described as follows:

> Employment protection is described along 21 basic items which can be classified in three main areas: (i) protection of regular workers against individual dismissal; (ii) regulation of temporary norms of employment; and (iii) specific requirements for collective dismissals. The information refers to employment protection provided through legislation and as a result of enforcement processes.[10]

Raw data have been converted to a 0–6 scale in which 0 is the lowest or weakest level of protection and higher values indicate more safeguards. Table 14-6 shows

TABLE 14-6 Results of Regression of Inequality on Union Density and Labor Protection

	Simple Regression			Multiple Regression		
	Gini ~ Union			Gini ~ Union + Labor		
	Estimate	Standard error	*t* statistic	Estimate	Standard error	*t* statistic
Constant	36.44	1.56	23.41***	40.58	2.30	17.62***
Union density	–0.14	0.04	–3.51**	–0.12	0.04	–3.38**
Labor protections	—	—	—	–2.24	0.99	–2.27*
	$R^2 = .39$			$R^2 = .53$		

$N = 21$. *significant at .05, **significant at .01, ***significant at .001.

Calculations carried out with more significant digits than reported; hence, quotients do not round exactly.

Source: Table 11-1 and Organisation for Economic Co-operation and Development, "OECD Indicators of Employment Protection." Accessed March 3, 2011. Available at http://www.oecd.org/document/11/0,3746,en_2649_37457_42695243_ 1_1_1_37457,00 .html

[10] OECD Indicators of Employment Protection. Accessed January 10, 2001. Available at http://www.oecd.org/document/11/0,3746,en_2649_37457_42695243_1_1_1_37457,00.html

what happens when this additional variable is included. (The table, by the way, follows a generally accepted way of summarizing the results.)

The left-hand panel shows the bivariate regression of Gini scores on union density; the right side of the table shows what happens when "labor protections" is added to the equation. There are several points to note.

- Now that a second independent variable has been added, the estimate of the constant *and* the regression coefficient for union density have changed slightly. That's because the extra variable has been factored in.
- As assessed by R^2, the fit of the model is greatly improved with the extra variable: "explained variation" increases from 39 percent to 53 percent.
- The negative signs attached to the coefficients have substantive meaning and support our general hypothesis: the greater working-class organizational and political strength, the *less* unequal the distribution of a nation's wealth.
- The regression coefficients are measured in units of the dependent variable, but their numerical magnitude reflects the measurement scales of the *X*s. Thus, the sizes of the coefficients are *not* directly comparable. Just because the beta for labor protections is –2.24 while the one for unions is –.12 does not necessarily mean that the former is a more important explanatory factor.
- All of the coefficients in the expanded model are statistically significant. (We describe hypothesis testing shortly.) This suggests that the two independent variables work partly independently to explain variation in *Y*. Presumably, if one of the variables were superfluous in the presence of the other, its coefficient value would be close to nil and would not be significant.
- In regression analysis, the effect of one independent variable is not simply added to the effect of another independent variable to get the "total" effect on *Y*, unless their covariation is zero; that is, the independent variables are independent. In this example, union concentration and labor protections are themselves weakly correlated ($r_{\text{union,labor}} = .19$). Regression-computing algorithms automatically adjust for this relationship, and the adjustment affects the magnitude of the coefficients and their standard errors.

Indeed, this last point again takes us to the meaning of regression analysis. The partial regression coefficient for union density is –.12, which means that inequality declines .12 units for every 1 percent increase in union density, *after* the labor protection variable has been held constant. The same is true for labor protection: a 1-unit increase in it brings a 2.24 *decrease* in inequality. And again, since the independent variables have different scales, we have to explore their construction to see what a "one-unit" change means in the real world.

Examining Residuals

We have emphasized that one goal of regression analysis is to make predictions. You may recall from the last chapter that the difference between predicted and observed values is called a *residual*. Although a formal analysis of residuals in the multivariate context can be tricky, a systematic scrutiny of their sizes may reveal some aspect of the data worth exploring further. Table 14-7 contains the predicted and observed Gini scores based on the most acceptable model we have found so far: $\hat{Y}_i = 40.58 - 0.12\text{union} - 2.24$ labor. The twelfth case, Japan, has been highlighted because its residual stands out for being (in absolute value) nearly twice as large as

TABLE 14-7 Observed and Predicted Gini Scores and Residuals

Country	Observed Gini	Predicted Gini	Residual ($Y_i - \hat{Y}_i$)
Australia	35.2	34.558	0.642
Austria	29.1	31.095	–1.995
Belgium	33.0	32.205	0.795
Canada	32.6	32.734	–0.134
Denmark	24.7	26.479	–1.779
Finland	26.9	25.160	1.740
France	32.7	34.113	–1.413
Germany	28.3	31.487	–3.187
Greece	34.3	33.672	0.628
Ireland	34.3	33.295	1.005
Italy	36.0	31.543	4.457
Japan	**24.9**	**33.930**	**–9.030**
Luxembourg	30.8	26.217	4.583
The Netherlands	30.9	32.636	–1.736
New Zealand	36.2	34.895	1.305
Norway	25.8	28.005	–2.205
Spain	34.7	32.269	2.431
Sweden	25.0	26.546	–1.546
Switzerland	33.7	33.578	0.122
UK	36.0	34.515	1.485
USA	40.8	36.968	3.832

any other residual in the table. For whatever reason, this case does not seem to fit the mold, and one wonders what would happen if it were (temporarily) eliminated.

To see what happens, we regress Gini on union density and labor protections for all countries except Japan and compare the results to the original model. Table 14-8 shows the comparison.

TABLE 14-8 The Effects of Deleting a Case with a Large Residual

Dataset	Estimated Model	Fit Indicator
Complete data ($N = 21$)	$\hat{Y}_i = 40.58^{***} - 0.12^{**}\text{union} - 2.24^{*}\text{Labor}$	$R^2 = .53$
Japan deleted ($N = 20$)	$\hat{Y}_i = 42.10^{***} - 0.14^{**}\text{union} - 2.48^{*}\text{Labor}$	$R^2 = .73$

*** = prob <.001, ** = prob < .01, * = prob <.05

As we might have anticipated, removing a case with such a large residual (Japan) greatly improves the apparent fit of the model, especially as measured by R^2 and the increases in significance. *Consequently, from now on we exclude Japan from our analyses.* We would *not* adopt this tactic in real research, except with solid statistical and substantive judgment to back up the decision, but in this chapter we only intend to explain regression and putting the "aberrant" case aside helps simplify the presentation. So far, we have found a linear model that seems to fit the data. The level of working-class mobilization does seem related to inequality. Hacker and Pierson's "Winner-Take-All Politics" article discussed in chapter 1 and elsewhere argues that growth in business power explains increases in inequality in the United States.[11]

Statistical Tests

Now that we have discussed regression coefficients, we can move on to testing hypotheses about them. Remember that (partial) regression coefficients are estimators of population parameters. Just because we have an estimated value of 2.24 (see table 14-8) for the partial coefficient for labor based on a sample of 21 cases, we cannot yet assert that this value or something close to it represents the true coefficient. Hence, hypothesis testing. A test for statistical significance, remember, requires the researcher to define a parameter(s) of interest; state null and alternative

[11] Jacob S. Hacker and Paul Pierson, "Winner-Take-All Politics: Public Policy, Political Organization, and the Precipitous Rise of Top Incomes in the United States," *Politics & Society* 38, no. 2 (2010): 152–204.

hypotheses; identify an appropriate sample estimator of the unknown parameter and determine its sample distribution under the null hypothesis and for the given sample size, *N*; establish a critical region for deciding when a sample outcome is unlikely if the null hypothesis is true; obtain the estimator and its standard error to calculate the observed test statistic; compare that against the critical value; decide whether or not to reject the null hypothesis; and interpret the results. (For testing purposes, we assume that the errors are normally distributed.)

There are two very closely related methods for testing statistical hypotheses: tests of individual coefficients and global or overall model tests.

INDIVIDUAL COEFFICIENTS. It turns out that under the assumptions of the regression model, the sampling distributions of the betas is known and well tabulated. For small samples, we use the *t* distribution; as the sample grows past 30 or 40, the *t* distribution increasingly approximates the standard normal. A general practice is to compute *t* statistics for each coefficient and compare the observed value with a critical *t* (or *z*) based on $N - K - 1$ degrees of freedom.[12] The observed *t* values are calculated, as shown in chapter 12, from the formula

$$t_{\text{observed}} = \frac{(\hat{\beta} - 0)}{\hat{\sigma}_{\hat{\beta}}} = \frac{\hat{\beta}}{\hat{\sigma}_{\hat{\beta}}},$$

where $\hat{\beta}$ is the estimated coefficient and $\hat{\sigma}_{\hat{\beta}}$ is the estimated standard error or standard deviation of the regression coefficient. Since it is standard practice these days to report standard errors along with the estimates themselves, if you have an estimate and its standard error, you can immediately calculate the observed *t*. If you also know the sample size, you can check its significance. We use zero in the numerator because in most published research, the null hypothesis is that the population coefficient, β, is zero. But in theory, you could check that a coefficient equals any hypothesized value.

Examine table 14-6. Take the coefficient and standard error for the partial regression of Gini on union while holding labor protections constant: –.12 and .04. Dividing gives an observed *t* of approximately –3.38. Since the regression is based on 21 cases and there are 2 independent variables in the model, union and labor, the degrees of freedom is 21 – 2 – 1 = 18. If you look in appendix B at the row for 18 degrees of freedom, you will see that the critical *t* (two-tailed test) at the

[12] The usual explanation for this formula for degrees of freedom is that to estimate the necessary standard deviations, we "lose" one degree of freedom for each regression coefficient plus one for the constant. A more precise explanation can be found in most statistics texts, such as Alan Agresti and Barbara Finlay, *Statistics for the Social Sciences*, 3rd ed. (Englewood Cliffs, N.J.: Prentice-Hall, 1997).

.005 level is 3.197, and at the .002 level it is 3.610. Thus, the probability under the null hypothesis of a coefficient this large or larger is somewhere between .002 and .005.

GLOBAL TEST. A global test assesses the overall model. In particular, the null hypothesis is

$$H_0: \beta_0 = \beta_1 = \beta_2 = \beta_3 = \ldots = \beta_K = 0.$$

That is, the test is of the hypothesis that all the coefficients (that is, the βs) equal zero. What about the rival or alternative hypothesis? For now, it is simply that *at least one* of the coefficients is nonzero in the population, but the particular one(s) is left unspecified. The mechanics of the test are a bit beyond the scope of the book, but we can sketch out the general idea. As in analysis of variance and two-variable regression, we calculate three sums of squares: total, explained (by the regression model), and unexplained or error. When a sum of squares is divided by its appropriate degrees of freedom, it becomes a "mean square," which is an estimator of a population variance (hence, the name "analysis of variance"). Under the null hypothesis that the regression parameters are all zero, the expected (long-run) values of the error and regression mean squares will be equal. If, on the other hand, the null hypothesis does not hold in some respect, the expected mean square for regression will be larger than the corresponding error mean square. This suggests that taking the ratio of the two would provide a way to judge how tenable the null hypothesis is. For given that H_0 is in fact true, then the expected value of this ratio will be 1.0.

The sums of squares are generated by most regression software. The information is usually arranged in the form of an ANOVA table like the ones we visited in the last chapter. Table 14-9 shows the results for the inequality data.

To deconstruct the table, look at the "Source" column. It lists the origins of the "explained" components—the contribution of union and labor—as well as the error and total, and next to them are the sums of squares. The total sum of squares (at the bottom) quantifies the total variation in the dependent variable, and you can see that it consists of three components, one for each of the independent variables and one for the error or residuals. The explained sum of squares by regression is about 73 percent of the total; this is the meaning of the multiple correlation coefficient shown in the last row. Two variables, union density and labor protection laws, account for more than half of the variation in Gini scores.

R^2 is a descriptive measure that shows us how well the model fits the data. But it is not in and of itself a hypothesis test. So go to the second column: it gives the sum of squares. The third column contains the degrees of freedom associated with each sum of squares. For the regression or explained portion, there is a degree of

TABLE 14-9 **Global Test**

Source	Sum of Squares	*df*	Mean Square	F_{Observed}
Explained				
Union	191.77	1	191.77	34.74***
Labor protection	65.77	1	65.77	12.10***
Union + Labor	257.55	2	128.77	23.33***
Unexplained				
Error (residual)	93.82	17	5.52	
Total	351.37	19	—	

$\text{Global}_{F2,17} = (257.55/2)/(93.82/17) = 23.33$***.

$R^2 = 257.55/351.37 = .73$.

Critical *F* with 2 and 17 degrees of freedom: .01 level = 6.11; .001 level = 10.66.

***Significant at .001.

freedom for each independent variable in the model; for the error, it is N minus the number of parameters including the constant, or $N - K - 1 = 20 - 2 - 1 = 17$, because the model contains a constant and two regression coefficients. The total sum of squares has $N - 1$ degrees of freedom. Notice that sums of squares and degrees of freedom are additive. For example, 191.77 + 65.77 + 93.82 = 351.37 and 2 + 17 = 19.

In the fourth column are the mean squares, which, as we said earlier, *independently* estimate the error variance if the null hypothesis is true. If the hypothesis does not hold, then the expected value of the regression mean squares will be larger than the error mean square. The table provides information for testing the coefficients one by one or as a group (a global test). So, for example, we need to compute the mean square errors and take their ratio. In the case of a single variable, such as union, the test statistic has the form

$$F_{\text{obs}(1,17)} = {(\text{Mean square for union})}\Big/{(\text{Mean square for error})} = \left(\frac{191.77}{1}\right)\Big/\left(\frac{93.82}{17}\right) = \frac{191.77}{5.52} = 34.74.$$

Under the null hypothesis, this ratio, called the F statistic, has a distribution like other statistics we have come across. As with the t and chi-square distributions, F's distribution is a family, each member of which is defined by the degrees of freedom used in the calculation of the two mean squares. So this statistic can be compared to a critical value obtained from appendix D. To do so, first decide on a level of significance (.05, .01, .001); then determine the "numerator" and "denominator" degrees of freedom. These are simply the quantities used to calculate the mean squares, the former being K, the number of variables in the model, and the latter being $N - K - 1$. With 1 and 17 degrees of freedom, the critical values at the .01 and .001 levels are, respectively, 8.40 and 15.72. Our observed value, 34.74, greatly exceeds the second, and we conclude that the partial regression coefficient is significant at the .001 level.

Alternatively, we can conduct an overall or global test by combining the regression sums of squares and degrees of freedom, as shown in the "Union + Labor" row. There on the right you will find the observed F for the model as a whole (i.e., Y as a linear function of the two independent variables). It is calculated the same way: find the total sum of squares due to regression, divide by the combined degrees of freedom, and divide that quantity by the error mean square:

$$F_{\text{obs}\,(2,17)} = {}^{\left(\text{Mean square for regression}\right)}\Big/_{\left(\text{Mean square for error}\right)} = {}^{\left({}^{257.55}/_{2}\right)}\Big/_{5.52} = {}^{128.77}/_{5.52} = -23.33.$$

What does statistical significance in this context tell us? We have rejected the null hypothesis that *all* the regression coefficients are zero. Naturally, that means that one or both independent variables is correlated with inequality even after controlling for the other. Which one(s)? Referring to the tests of the individual coefficients in the table 4-9, we see that both union concentration and labor protections are significant at the .001 level.

If you look in appendix D, you will see that the critical F with 2 and 17 degrees of freedom at the .001 level of significance is 10.66. Our estimate barely misses that standard, so we say the model is significant at the .001 level.

COMPARISON OF NESTED MODELS. The analysis of linear models can be looked at still another way. Suppose we add a third independent variable to the analysis of Gini data and estimate this model:

$$\text{Full:}\quad \hat{Y} = \hat{\beta}_0 + \hat{\beta}_1\text{Union} + \hat{\beta}_2\text{Labor} + \hat{\beta}_3\text{Employ}.$$

Employ stands for "employment ratio," which is the proportion of a nation's working-age population actually in the labor force. This contrasts with the previous, two-independent variable model:

$$\text{Reduced:}\quad \hat{Y} = \hat{\beta}_0 + \hat{\beta}_1\text{Union} + \hat{\beta}_2\text{Labor},$$

which, as can be seen, is "nested" within the larger one (Full) in the sense that all of its independent variables are a subset of those in Full. (There is no variable in Reduced that is not in the full model.) Since the complete model has more explanatory terms than its cousin, it has at least as much and presumably more explanatory power. The difference will show up in the R^2s and explained regression sum of squares, both of which in the full model will be equal to or greater than in the reduced model. How much greater? We could obtain sums of squares and degrees of freedom from the two models and insert them in an ANOVA table exactly as above. This procedure, however, boils down to an *incremental* F-*statistic:*

$$F_{K-p,N-K-1} = \frac{\left(SSRegress_{\text{Full}} - SSRegress_{\text{Reduced}}\right)/\left(K-p\right)}{SSRegress_{\text{Full}}/\left(N-k-1\right)} = \frac{\left(N-K-1\right)}{\left(K-p\right)} \times \frac{\left(R^2_{\text{Full}} - R^2_{\text{Reduced}}\right)}{\left(1-R^2_{\text{Full}}\right)}.$$

This quantity has an F distribution with $K-p$ degrees of freedom for the numerator and $N-K-1$ degrees of freedom in the denominator, where K and p are the number of independent variables in the full model and reduced models, respectively.

Calculating this statistic is straightforward with the appropriate software: obtain R^2 for the full and reduced models and determine how many variables are in each. Then plug them into the formula. By way of illustration, table 14-10 shows the two estimated regression models with their R^2s.

With these numbers in hand, we can assess the improvement in fit obtained from adding employment ratio to the mix: R^2 edges up from .53 to .55. This doesn't look like a big deal. But is it a statistically significant improvement? The test statistic,

$$F_{1,17} = \frac{\left(N-K-1\right)}{\left(K-p\right)}\frac{\left(R^2_{\text{Full}} - R^2_{\text{Reduced}}\right)}{\left(1-R^2_{\text{Full}}\right)} = \frac{\left(20-3-1\right)}{\left(3-2\right)}\frac{\left(.55-.53\right)}{\left(1-.55\right)} = .71,$$

tells us "no, not significant at even the .1 level." (The critical F with 1 and 17 degrees of freedom is 4.45 and the observed F is far less, so it does not fall in the

TABLE 14-10 **Full and Reduced Models with R^2**

Model	Estimated Equation	Number of Predictors	R^2
Full	$\hat{Y}_i = 44.36 - 0.13\text{union} - 2.45\text{labor} - 0.03\text{employ.}$	$K = 3$	.55
Reduced	$\hat{Y}_i = 42.09 - 0.14\text{union} - 2.48\text{labor.}$	$p = 2$	.53

$N = 20$, $F_{1,17} = .619$, prob > .44

critical region.) Thus, we do not reject the null hypothesis that $\beta_{\text{employ}} = 0$. In simple words, the extra variable adds nothing to the explanation of inequality. This "negative" finding may or may not have practical import. If the employment ratio variable loomed large in discussions of politics and inequality, we would spend time discussing possible reasons for the lack of significance. If, on the other hand, we had added it to the model just to see what would happen, we would probably mention but not dwell on the result.

The current example is trivial because we reduced the full model by just one variable and could have anticipated the finding from the small boost in R^2. But the general strategy of comparing models with an incremental test is quite flexible and handy. We'll see an illustration in the next section, but for now assume we have, say, five demographic variables and two economic indicators in a full model. We might want to evaluate the impact of dropping the first five as a group. The complete model would then have $K = 7$ parameters plus the constant, while the reduced one would have just two coefficients plus the constant. If R^2_{Reduced} is not much less than R^2_{Full} and if the F is insignificant, we might conclude that demographic factors are not essential to the analysis.

CONFIDENCE INTERVALS. After identifying significant and/or interesting partial regression coefficients, one can place confidence intervals around the estimated values. To review a confidence interval for a given alpha (level of significance), the equation has this form:

$$\text{Estimator} \pm t_{(1-\alpha)/2},\ \hat{\sigma}_{\text{Estimator}},$$

where $t_{(1-\alpha)/2,}$ is the critical value of a test statistic (usually t) at the $(1 - \alpha)/2$ level of significance with the appropriate degrees of freedom, and $\hat{\sigma}_{\text{Estimator}}$ is the standard error of the estimator. The estimator and its standard deviation fall out of the regression analysis, as we have already seen.

Earlier we presented a model that contained two independent variables, union concentration and labor protection. The estimated equation is

$$\begin{gathered}\hat{Y}_i = 42.10^{***} - 0.14 \text{ union}^{***} - 2.48 \text{ labor}^{**}.\\ (1.711) \quad (.026) \quad (0.719)\end{gathered}$$

(Note that, as usual, the stars indicate the level of significance.) If we want 99 percent intervals for each coefficient, the critical t with 17 degrees of freedom is 2.898. Putting all this in the formula gives, for example, the interval for union:

$$\begin{aligned}\text{CI}_{.99} &= -.14 \pm 2.898(.026)\\ &= -0.22 \text{ and} -0.06.\end{aligned}$$

TABLE 14-11 99 Percent Confidence Intervals for Inequality Model

Parameter	Estimate	Lower	Upper
Constant[a]	42.10,	37.14	47.06
Union	–.14	–.22	–.060
Labor protection	–2.48	–4.566	–.398

[a] The constant is usually not a major concern. We include it here for illustrative purposes.

Table 14-11 summarizes the results.

These confidence intervals agree with the significance shown in the model's equation: none of them includes zero. As we said in chapter 12, confidence limits provide another way of looking at hypothesis testing, so the fact that the limits exclude zero is just another way of saying that the null hypothesis is not accepted.

Categorical Variables and Linear Models

Suppose we have a categorical variable such as region or gender. How can it be entered into a regression model? It turns out to be surprisingly easy because there are various ways of doing so. One method that won't work (at least not usually, and not very well) is to treat any numbers assigned as group names as just plain numbers. If region labels are 1, 2, 3, . . . , but are used for convenience, they won't function as numbers in regression analysis. Therefore, we need a different approach.

A common method is dummy variable coding. A **dummy variable** has just *two* values: 0 for the presence (or absence) of a characteristic, group membership, condition, and so on, and 1 for its absence (or presence). The digits 0 and 1 are more or less arbitrary—we could use 1.5 and 100 to mark the presence and absence of a trait—but 0 and 1 lead to some facile interpretations. Dummy variables—sometimes we use the phrase *indicator variables*—are widely used to convert categorical data into a form suitable for numerical analysis.[13] Here is the general idea: Convert an ordinal or nominal variable, *X*, into a set of dummy variables, one for each category. The dummy variables are created by assigning the value 1 if an

[13] This is not the only way to treat categorical data. Another common procedure is "effect coding" or "deviation coding," which uses scores –1, 0, and 1 as measurement units. See Graeme Hutcheson and Nick Sofroniou, *The Multivariate Social Scientist* (Thousand Oaks, Calif.: Sage, 1999), 85–94.

observation is a member of that category and 0 otherwise. If a variable has $J = 4$ classes, any individual will get a score of 1 on one of the dummy variables and 0 on the other three.

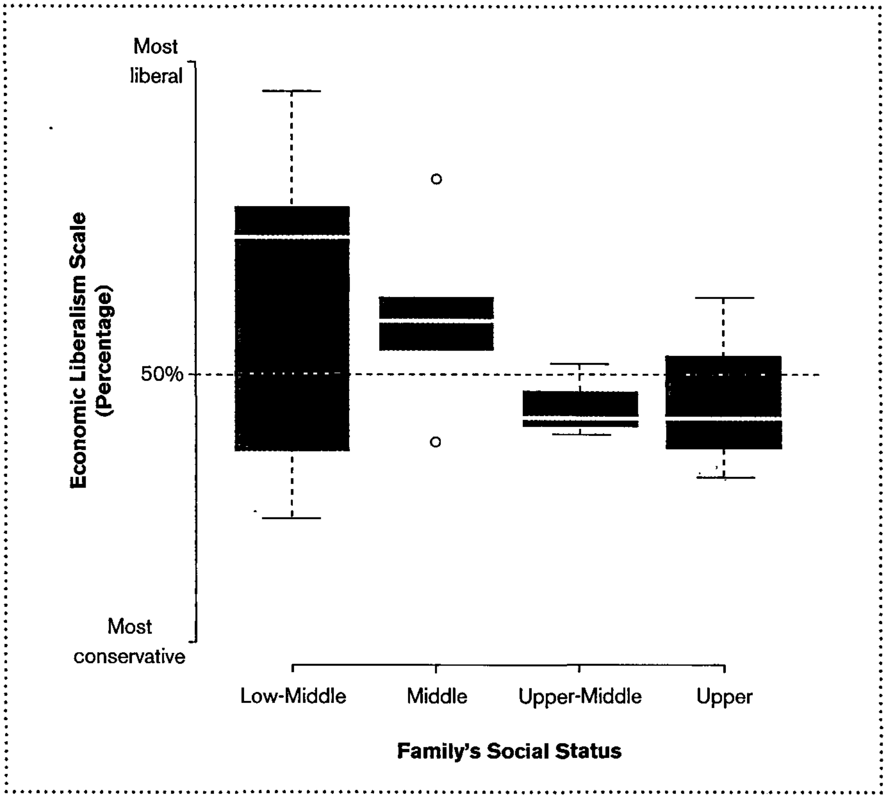
As always, a concrete example helps. Return once again to judicial decision making. Following journalists and scholars, we advanced the proposition that Supreme Court justices do not (no doubt cannot) banish all political preferences or predispositions from their minds when they take office. Instead, we propose, justices carry those predispositions with them into their deliberations. To test this idea, we compared the justices' liberalism-conservatism indicator on various issues to the party and attitudes of the president and Senate that nominated and confirmed them. Although the universe is quite small (twenty-three justices), we found modest associations between partisanship and their rulings in a number of policy areas. Let's explore the idea more deeply by measuring the impact (if any!) of the justices' family social status when growing up. The dataset we have been using—US Supreme Court Justices Database—contains a variable that "indicates the general socioeconomic status of the nominee's family during his or her childhood." The Court members are assigned to one of five categories: lower, lower-middle, middle, upper-middle, and upper. Because there are so few cases in the first group, we combined lower and lower-middle into a four-category scheme.

In order to follow what comes later, look at figure 14-4. Here is another boxplot that compares the distribution of economic liberalism scores (Y) within levels of family social status (X). We see that the medians (the solid lines in the boxes) trend downward as we move from lower to upper class. The substantive interpretation is that the lower the category—here the categories have an implicit order—the higher the economic liberalism measure and vice versa. Those justices born and raised in a particular milieu seem to carry their socialization into their adult lives. Needless to say, this conclusion is very tentative since it rests on a tiny sample and unverified assumptions about errors in the model. Still, let's analyze the data more formally, if for no other reason than just to provide a numerical example.

Back to dummy variables. With four socio-economic groups we need four dummy variables, one for each category of status. However, in the ensuing analysis, we have to drop one of the variables in order to make the model estimable. A quick example shows why. Denote the categorical variable Z and its individual category dummy variables as Z_j. Table 14-12 lists four justices and their families' socioeconomic background category. If you spend a little time looking at the table, you can tell that, if you know a person's score on any three variables, you can predict *exactly* his or her value on the remaining one.

Start with Earl Warren. Once you know his first dummy variable score is 1, a little thought shows that his scores on the other three must be zero. Why? Because a 1 indicates membership in a class and belonging to it automatically means he can't be in any of the others. Conversely, if you knew that Warren's scores on the last

FIGURE 14-4 "Liberalism" of Economic Rulings by Justice's Family Background



Source: Lee Epstein et al., US Supreme Court Justices Database, 2010.

TABLE 14-12 US Supreme Court Justices and Socioeconomic Background

		Dummy Variables (Z_i)			
Justice	Status category	$Z_{\text{Low-middle}}$	Z_{Middle}	$Z_{\text{Middle-upper}}$	Z_{Upper}
Earl Warren	Low-middle	1	0	0	0
Byron White	Middle	0	1	0	0
Sandra Day O'Connor	Upper-middle	0	0	1	0
John Paul Stevens	Upper	0	0	0	1

three are 0, you wouldn't even have to look to know that he gets a 1 on the first variable. In statistical language, $Z_{\text{Low-middle}}$, Z_{Middle}, $Z_{\text{Middle-upper}}$, and Z_{Upper} are perfectly linearly related. As the list of regression requirements stated, this condition is a no-no. A peek under the hood of the computing machinery would show you that there are too many coefficients to estimate given the available information. The way to avoid this so-called multicolinearity is to drop one of the dummy variables from the analysis. The principle is this: *for a variable with J categories, create J – 1 indicator variables.*

Dummy variables are often defined in these terms:

$$Z_j = 1, \text{ if observation is in category } j;\ Z_j = 0 \text{ otherwise.}$$

Since there are four statuses, we need 4 – 1 = 3 dummy variables. We convert status (Z) into three indicator variables as follows:

Middle:	Z_2	= 1, if a justice is from middle-class background.
		= 0, otherwise.
Middle-upper:	Z_3	= 1, if a justice is from upper-middle class background.
		= 0, otherwise.
Upper	Z_4	= 1, if a justice is from upper-class background.
		= 0, otherwise.

There is no Z_1 for low-middle. Instead, it will serve as a reference or base or comparison category against which the effects of X are measured. Any category can serve as the reference point, but always try to pick one with substantive meaning because you'll be saying things like "Compared to justices in the lowest status, the effect of 'moving' to the next higher level is such and such." If you picked a category in the middle of the scale, comparisons might be a tad harder to sort out.

Whatever the choice, these variables are treated as numeric and inserted in the model like any other set of independent variables. To accommodate different kinds of variables, we will for convenience use lowercase lambda (λ) and lowercase delta (δ) to stand for the population partial regression coefficients for dummy variables.

And we now denote the constant with alpha (α). These symbols are used for different purposes elsewhere, but the context should make their meaning clear.

The regression model for dummy variables take various forms depending on the kind and number of variables in the model (see table 14-13). Please don't turn away from the equations. They are not as intractable as they seem—we'll discuss the details in a minute.

TABLE 14-13 **Some Models Using Dummy Variables**

Procedure	Variables in Model	Model Equation	Comment
Single dummy variable regression	Dependent variable, *Y*: quantitative One categorical independent variable (*Z*) with *J* classes	$\hat{Y} = \hat{\alpha} + \hat{\lambda}_2 Z_2 + \hat{\lambda}_3 Z_3 + \ldots + \lambda_J Z_K.$	Equivalent to ANOVA
Two-variable dummy variable regression	Dependent variable, *Y*: quantitative Two categorical independent variables (*Z* and *W*) with *K* and *J* levels, respectively	$\hat{Y} = \hat{\alpha} + \hat{\lambda}_2 Z_2 + \hat{\lambda}_3 Z_3 \ldots + \hat{\lambda}_K Z_K + \hat{\delta}_2 W_2 + \hat{\delta}_2 W_3 + \ldots + \hat{\delta}_J W_J.$	Additional categorical variables can be added. Possible alternative to two-way ANOVA.
Analysis of covariance	Dependent variable, *Y*: quantitative Quantitative X_1 and categorized independent variable (Z)	$\hat{Y} = \hat{\alpha} + \hat{\beta}_1 X_1 + \hat{\lambda}_2 Z_2 + \hat{\lambda}_3 Z_3 + \ldots$	β constant across groups
Analysis of covariance with interaction	Dependent variable, *Y*: quantitative Quantitative and categorical independent variables plus interaction term: $I = XZ$	$\hat{Y} = \hat{\alpha} + \hat{\beta}_1 X_1 + \hat{\lambda}_2 Z_2 + \hat{\lambda}_3 Z_3 + \ldots + \hat{\omega}_1 XZ_2 + \hat{\omega}_2 XZ_3$ $= \hat{\alpha} + \hat{\beta}_1 X_1 + \hat{\lambda}_2 Z_2 + \hat{\lambda}_3 Z_3 + \ldots + \hat{\omega}_1 I_1 + \hat{\omega}_2 I_2 \ldots$	βs across groups are *not* equal.

Interaction: $I = XZ$.

α is the constant; λ and δ are regression coefficients for categorical variables *Z* and *W*, respectively; β is the regression coefficient for *X*; and ω is the regression coefficient for the interaction variable, *I*.

Note: The *first* category is the base or comparison point and is omitted from the models. Any category, however, can serve this purpose. Make the choice as substantively meaningful as possible.

In the current example, we have just one categorical variable (status) with four levels that we want to utilize as a predictor. In symbolic form the model is

$$\hat{Y}_i = \hat{\alpha} + \hat{\lambda}_2 Z_2 + \hat{\lambda}_3 Z_3 + \hat{\lambda}_4 Z_4,$$

and the estimation turns out to be

$$\hat{Y}_{\text{Liberalism}} = 57.17 - 1.06\text{Middle} - 11.21\text{Upper-middle} - 10.47\text{Upper}.$$

The complete results appear in table 14-14. We'll discuss the statistical tests later; for now, concentrate on the meaning of the coefficients.

In the case of these types of variables, one can apply the standard rote interpretation to the estimated parameters: "a one-unit change in . . .". This is perfectly valid and will make sense if one reflects on what a "one-unit change" in a dummy variable would mean. Thinking abstractly, one could imagine a justice somehow coming from a different social and economic environment. What would the "effect" be? The regression parameter gives the answer. Changing Earl Warren's family status from low-middle to middle would be expected to lead to a 1 percent (1.06) *decrease* in his economic liberalism score.

TABLE 14-14 Supreme Court Decisions by Social Status

Category	Coefficient Estimate	Standard Error	Observed *t*
Constant = mean of observations for base category	57.173	4.76	12.01**
Middle	–1.06	7.01	–0.15
Middle-upper	–11.21	7.90	–1.39
Upper	–10.47	7.90	–1.33

ANOVA Table:

Source	Sum of squares	*df*	Mean squares	F_{Obs}
Status (Z_j)	535.57	3	178.53	1.1245
Error	2698.90	17	158.76	
Total	3234.47	20		

Critical *F* with 3 and 17 degrees of freedom = 3.20 at the .05 level.

Or we can write out the equations for each category. Doing so takes advantage of the fact that some coefficients will drop out because if an observation does not belong in a group, its score on the corresponding dummy variables is zero. For instance, look first at only those justices in the comparison category, low-middle. Substituting values for the dummy variables in the equation gives

$$\hat{Y}_{\text{Liberalism}} = 57.17 - 1.06(0) - 11.21(0) - 10.47(0) = 57.17$$

All these people are in the first category and so have zero values on the middle, middle-upper, and upper dummy variables. This result, 57.16 percent, is the predicted or expected liberalism for those appointees from "humble" origins. (It also equals the mean economic liberalism of the justices with lower- to middle-class origins, as we pointed out when explaining the meaning of the regression constant.) But what about someone from a middle-class family, the next category? Just write out the equation to find out:

$$\hat{Y}_{\text{Middle}} = 57.17 - 1.06(1) - 11.21(0) - 10.47(0) = 57.16 + (-1.06) = 56.11.$$

The predicted value has dropped a modest 1.06 percent. We can keep going and derive some meaning from the remaining coefficients by making the appropriate substitutions:

$$\hat{Y}_{\text{Middle-upper}} = 57.17 - 1.06(0) - 11.21(1) - 10.47(0) = 57.17 - 11.21 = 45.96.$$
$$\hat{Y}_{\text{Upper}} = 57.17 - 1.06(0) - 11.21(0) - 10.47(1) = 57.16 + (-10.47) = 46.70.$$

The mean economic percentage drops a precipitous 11 percent when moving from low-middle to middle-upper and 10.5 percent when moving into the upper status category. We can simply read these effects from the estimates in the table. (See table 14-14.) In the one categorical variable case, the regression constant equals the mean *Y* for those in the reference or comparison class. The regression coefficient measures the "effect" of moving from one category to the next. For instance, to find the mean liberalism of middle-class justices, just add the regression coefficient to the constant: 57.17 + (–1.06) = 56.11. That is, the consequence of a move from low-middle to middle is a lowering of the average liberalism by about 1 percent. So, in this simple case, the measurement of the "effect" of middle-class status is 1.06 percent.

Does a Model Fit? ANOVA and Dummy Variable Regression

Did you notice that in the previous example, we were effectively comparing one mean with another? As you might recall, that is what analysis of variance does:

it tests for differences in group means. And so does regression analysis. Both are essentially efforts to see how independent variables explain variation in Y. Consequently, dummy variable regression gives the same results as ANOVA.

This point becomes clearer when we come to hypothesis tests. For now, look again at table 14-14. The bottom of the table displays the ANOVA analysis. It answers the general question: Does knowledge of a justice's social background help predict the direction and content of his or her rulings on economic disputes? More formally, it compares a full model $(\hat{Y}_i = \hat{a} + \hat{\lambda}_2 Z_2 + \hat{\lambda}_3 Z_3 + \hat{\lambda}_4 Z_4)_i$ with a reduced model $(\hat{Y}_i = \alpha)$. The latter model effectively states that there is no explained variation and so $R^2_{\text{Reduced}} = 0$. Let's use the incremental F-test introduced above. For that we need the following:

- $N = 21$, the sample size (number of justices excluding recent appointments)
- $K = 3$, the number of parameters (*excluding* the constant) in the full model
- $p = 0$, the number of parameters (*excluding* the constant) in the reduced model
- $R^2_{\text{Full}} = 535.57/3234.47 = .1656$ (sums of squares are in table 14-14)
- $R^2_{\text{Reduced}} = 0/3234.47 = 0$

Inserting these values in the formula gives

$$F_{1,17} = \frac{(N - K - 1)}{(K - p)} \frac{\left(R^2_{\text{Full}} - R^2_{\text{Reduced}}\right)}{\left(1 - R^2_{\text{Full}}\right)} = \frac{(21 - 3 - 1)}{(3 - 0)} \frac{(.1656 - 0)}{(1 - .1652)} \approx 1.245,$$

which fails to exceed the critical value at even the .10 level.

Like all statistical analyses, t- and F-tests rest on assumptions about the distribution of the errors. We generally prefer using the incremental test over calculating multiple individual t statistics, but both are widely reported in the literature.

Models with Quantitative and Categorical Variables: Interaction and Analysis of Covariance (ANCOVA)

There is no reason we cannot simultaneously investigate the effects of both quantitative and categorical variables on a numeric dependent variable. As a matter of fact, doing so often leads to interesting conclusions. Here is a continuation of a previous example, inequality in postindustrial democracies. We are trying to find out what factors explain cross-national variation in inequality. We started with union density, then added labor protection to obtain a model that fits reasonably well.

Can it be improved further? Here is where literature review comes to the fore. A survey of books and articles reveals that some scholars believe there is a difference between European and Anglo-American political culture when it comes to attitudes about workers' rights, social insurance, and welfare spending, all of which affect the distribution of wealth. To test this proposition we created a crude indicator:

$$\text{Culture } (Z_2) = 1 \text{ if Anglo-American, } Z_2 = 0 \text{ otherwise.}$$

(The excluded variable is $Z_1 = 1$ if European, 0 otherwise. So our reference category is Europe, and we will gauge the effects on inequality of "changing" from it to Anglo-American.)

The general form of the equation is

$$\hat{Y}_i = \hat{\alpha}_0 + \hat{\beta}_{\text{Union}}\text{Union} + \hat{\lambda}_{\text{Culture}}\text{Culture}.$$

The first coefficient is, of course, the constant, and the second is the partial regression coefficient of inequality on union membership with "culture" held constant. The coefficient ($\hat{\beta}_{\text{Culture}}$) shows the effects of "moving" from one culture to another. The estimated equation is

$$\begin{aligned}\hat{Y}_{\text{Gini}} = 35&.42 - .13\text{Union} + 3.60\text{Culture}.\\ (1&.37) \quad (.03) \quad (1.32)\end{aligned}$$

Following standard practice, we present the estimates along with their standard errors. If you divide one into the other, you can obtain observed *t* statistics to decide which will be statistically significant at a particular alpha level. Keep in mind that due to the small *N* and the fact that basic assumptions (e.g., normally distributed errors) may not be met, we should take the hypothesis tests with a grain of salt. Besides, we'll come back to them later when describing inference for multiple regression. More important at the moment is nailing down what these numbers mean. Once more, writing out the estimated equation and substituting various values for the independent variable help. First, what happens if there are *no* unions (union = 0) and we are looking at only European countries (culture = 0)? The equation simplifies greatly:

$$\hat{Y}_{\text{Gini}} = 35.92 - .13(0) + 2.85(0) = 35.92.$$

This is the mean Gini score for those nations meeting these conditions (union = culture = 0) and appears meaningless because no country is entirely without a labor movement. But it provides a baseline for comparison. If a culture could somehow switch from European to Anglo-American, the effect would be to increase inequality:

$$\hat{Y}_{\text{Gini}} = 35.92 - .13(0) + 2.85(1) = 35.92 + 2.85 = 38.77.$$

That is, inequality would increase to 39. (We know there should be an increase because of the plus sign attached to the coefficient.) In words, the Anglo-American nations are a bit more unequal (by this standard) than those on the Continent. But our real objective is to see how the two independent variables work together to achieve their effects. Let's set union density at its mean for the dataset (excluding Japan), 35.51, and again consider the European nations (culture = 0). Substituting these values into the estimated model, we get

$$\hat{Y}_{\text{Gini}} = 35.92 - .13(35.51) + 2.85(0) = 35.92 - 4.62 = 31.30.$$

In other words, the predicted Gini score for European nations with an average level of unionization is 31.30. How do non-European countries (culture = 1) with the same mean union density stack up? Just plug the data values into the equation:

$$\hat{Y}_{\text{Gini}} = 35.92 - .13(35.51) + 2.85(1) = 35.92 - 4.62 + 2.85 = 34.15.$$

We see that inequality increases: Europe has slightly more economic equality than Anglo-American nations do *even when the level of unionization is controlled*. That is, culture adds a bit to our understanding of political economy over and above what social and economic factors supply. (Needless to say, "culture" is a broad-brush attempt to capture complex and nuanced social, economic, and political aspects of society, and we employ it mainly for expository reasons.)

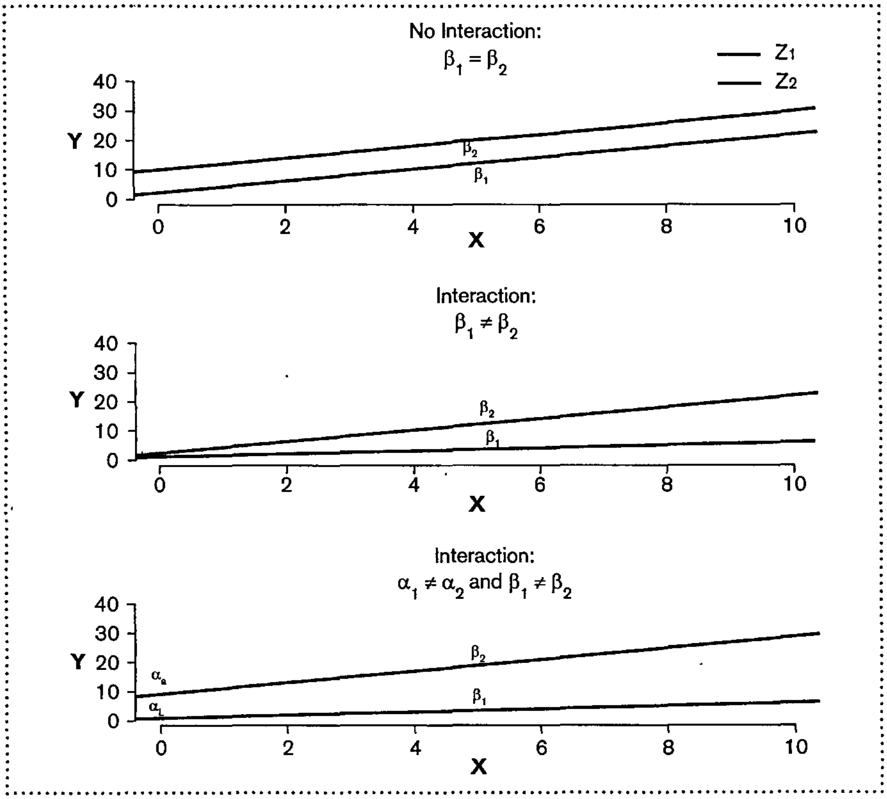
We have been following a general method of inserting prespecified and meaningful values into estimation equations. It's a trick that is helpful for understanding the next application of dummy variables to regression analysis.

Interaction

Earlier in the chapter, we introduced a concept that is very important in linear models, interaction. Interaction, as we said then, means that the nature of a relationship between two variables, *Y* and *X*, depends on levels of a third variable, *Z*. The test is: Does holding *Z* constant—that is, measuring the *Y*-*X* relationship at each level or value (or interval of values) of *Z*—affect the relationship between *Y* and *X*? Or, to put it another way, looking to see if the relationship between *Y* and *X* is different for different values of *Z*. Testing and measuring interaction effects in the context of linear models are often called the analysis of covariance (ANCOVA). In the simplest case—the one we present here—there is a quantitative dependent variable, *Y*; a quantitative independent variable, *X*; plus a categorical factor with *J* categories.

Table 14-5(b) illustrated the idea with categorical variables. Figure 14-5 does the same for interaction between two quantitative variables, *X* and *Y*, and one qualitative variable with two levels, Z_1 and Z_2. (For convenience, we denote them as "group 1" and "group 2.") The figure contains two panels. The first illustrates "no interaction." It is meant to show a situation in which the effects of *X* on *Y* are the same regardless of the value.of *Z*. The regression constants ($\alpha_2 > \alpha_1$) differ so that for a given value of *X*, observations in group 2 have higher *Y* values than do those in the first level. But the difference is constant across the range of *X*. Even more important, the nature of the relationship between *Y* and *X*—its strength and direction—are the same in both groups. Thus, *X* has the same impact on *Y* no matter what *Z* is. Not so in the second graph.

FIGURE 14-5 Interaction



Here we see that the two slopes differ: $\beta_2 > \beta_1$. This suggests that for a one-unit change in X, Y increases more in group 2 than group 1. Depending on the context, the difference could have theoretical or substantive importance. Seen another way, note that, although the two regression constants are the same, the difference between the lines is not constant. The third plot is a picture of a situation in which interaction exists ($\beta_2 > \beta_1$) and the intercepts also differ ($\alpha_2 > \alpha_1$). In cases where this condition holds, we might say that there are two separate linear processes at work: one that applies to group 1 and another that applies to group 2.

No real dataset will follow exactly these patterns, but plots and regression analysis with dummy variables will indicate when interaction might and might not be present. An interaction "variable" is created from the independent and categorical variables by simple multiplication:

$$\text{interaction: } I = XZ.$$

How do you multiply a categorical variable by anything? You don't. Instead, you multiply X by each dummy variable representing the categories of Z. So if Z has $J = 5$ categories, there will be 4 dummy variables, each of which is multiplied by X to create four interaction variables. So a model with an interaction term looks like this. Consider a Z with two categories, the first of which is treated as the reference point. It appears in the model as Z_2. The interaction is represented by multiplying the two explanatory variables in the equation:

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta}_1 X + \hat{\lambda}_2 Z_2 + \hat{\eta}_1 XZ_2 = \hat{\alpha} + \hat{\beta}_1 X + \hat{\lambda}_2 Z_2 + \hat{\eta}_i I.$$

Important: For various reasons related to the interpretation of parameters, it is usually advisable to retain *all* "parent" variables in the interaction term. (Technically, these are called "hierarchical" models.) Thus, if I consists of XZ, then the two variables should also appear separately in the equation. (For example, don't force a computer to estimate this model: $\hat{Y}_i = \hat{\alpha} + \hat{\eta}_1 I$, where $I = XZ$.)

Another way of looking at interaction is this: the independent variable's influence on Y consists of more than its additive main effect (represented by the "+ X" term) but is "supplemented" by an additional multiplicative effect of the form $I = XZ_j$. (The same can be said for Z.)

The interaction component can be confusing until one revisits the substitution method. Pick a set of meaningful values for the independent variables, substitute them into the estimated model, and observe how changes in one factor affect the response variable while the others are held constant. To demonstrate, we extend the previous model by adding an interaction term:

$$\begin{array}{llllll} \hat{Y}_{\text{Gini}} = & 35.15 & - .12\ \text{Union} & + 7.73\ \text{Culture} & - 0.16\ \text{Interaction.} \\ & (1.38) & (.03) & (4.00) & (.15) \end{array}$$

The "main" (marginal) effects of unionization and culture are –.13 and +8.27, respectively. (Note again, moving to Anglo-American culture decreases the Gini score by about 8 points, which means a decrease in inequality.) There is a single interaction term, –.20, which is the only new wrinkle. We need to see what it means. The best way for now is to make substitutions as we have been doing. So let's say we want the predicted inequality score for European countries whose union density percentage is 35.51, as we just saw. With $X = 35.51$, $Z_1 = 0$, and $I = 35.51 \times 0 = 0$, the equation reduces to

$$\begin{aligned}\hat{Y}_{\text{Gini}} &= 35.43 - .13(35.51) + 8.27(0) - .20(35.51\,X\,0)\\ &= 35.43 - 4.62\\ &= 30.81.\end{aligned}$$

Compare this to the prediction for the Anglo-American countries,

$$\begin{aligned}\hat{Y}_{\text{Gini}} &= 35.43 - .13(35.51) + 8.27(1) - .20(35.51\,X\,1)\\ &= (35.43 + 8.27) - (.13(35.51) - .20(35.51)) \text{ after rearrangement}\\ &= 43.7 + 35.51(-13 - .20) \text{ factor out } 35.51\\ &=31.98.\end{aligned}$$

Hence, we observe quite a jump in the predicted inequality (given $X = 35.51$), from about 31 to about 32. There seems to be an added boost over and above that predicted by the main effects of union and culture.

Standardized Regression Coefficients

As discussed in chapter 13, a regression coefficient calculated from standardized variables is called a standardized regression coefficient or, sometimes, a beta weight. Under certain, restricted circumstances, it might indicate the relative importance of each independent variable in explaining the variation in the dependent variable when all other variables are controlled for. Standardizing a variable, you may remember from chapter 13, means subtracting its mean from each individual value and dividing by the standard deviation. The results are frequently called *scaled* variables, a term we use intermittently hereafter. To obtain the standardized regression coefficients, you standardize all the variables, including Y, and then regress the standardized Y on the standardized Xs. It is the same procedure demonstrated in chapter 13, except that now there are more than two variables. A standardized coefficient shows the partial effects of an X on Y in standard deviation units. The larger the absolute value, the greater the effect of a one-standard-deviation change

HELPFUL HINTS

Interpreting Models with Dummy Variables and Interaction

If dummy variables are coded 0 and 1, one can simplify equations and gain a better understanding of what they mean by replacing *Z*s with these values and rearranging terms. Consequently, a model for *Y* with a quantitative variable, *X*, and a single categorical variable, *Z*, with $J = 3$ categories looks like this in general:

$$\hat{Y} = \hat{\alpha} + \hat{\beta}_X X + \hat{\lambda}_2 Z_2 + \hat{\lambda}_3 Z_3.$$

Z_1 has been omitted, and the first category serves as the reference point. Since the *Z*s can only be 0 or 1, we can simplify the equation. For members of the reference class ($Z_1 = 1$), Z_2 and Z_3 are both 0, and terms associated with them drop out:

$$\hat{Y} = \hat{\alpha} + \hat{\beta}_X X + \hat{\lambda}_2(0) + \hat{\lambda}_3(0) = \hat{\alpha} + \hat{\beta}_X X.$$

This looks like a simple two-variable regression, and indeed it is. But it is restricted to the observations in the base category. Now look what happens when we add the second group for which $Z_2 = 1$ and $Z_1 = Z_3 = 0$. Make the substitutions and rearrange terms:

$$\begin{aligned}\hat{Y} &= \hat{\alpha} + \hat{\beta}_X X + \hat{\lambda}_2(1) + \hat{\lambda}_3(0)\\ &= \hat{\alpha} + \hat{\beta}_X X + \hat{\lambda}_2\\ &= \left(\hat{\alpha} + \hat{\lambda}_2\right) + \hat{\beta}_X X\\ &= \hat{\alpha}' + \hat{\beta}_X X.\end{aligned}$$

This, too, appears to be a linear regression with the same main effect of *X* but a new regression constant,. As you can see, it consists of two parts, the original constant, $\hat{\alpha}'$, and the coefficient for category 2, $\hat{\lambda}_2$. Recalling that the constant can be interpreted as the expected value of *Y* when *X* is zero, we see that the regression coefficient for a dummy variable can be interpreted as an "adjustment" (up or down depending on its sign) to the expected value of the dependent variable.

Deconstruction of an interaction model follows the same logic:

$$\hat{Y} = \hat{\alpha} + \hat{\beta}_X X + \hat{\lambda}_2 Z_2 + \hat{\lambda}_3 Z_3 + \hat{\delta}_2 XZ_2 + \hat{\delta}_3 XZ_3.$$

Consider Z_1. Then, $Z_2 = Z_3 = 0$ and $I_2 = XZ_2 = I_3 = XZ_3 = 0$, and simplification follows:

$$\begin{aligned}\hat{Y} &= \hat{\alpha} + \hat{\beta}_X X + \hat{\lambda}_2(0) + \hat{\lambda}_3(0) + \hat{\delta}_2 X(0) + \hat{\delta}_3 X(0),\\ &= \hat{\alpha} + \hat{\beta}_X X.\end{aligned}$$

Another simple regression equation. And, as you might anticipate, the presence of interaction is going to affect both the regression constant and partial regression coefficient. To see this point, compare the previous equation with what we get when we insert $Z_2 = 1$ in the model and Z_1 and Z_3 are both zero. Notice that now $I_2 = XZ_2 = X$:

$$\begin{aligned}\hat{Y} &= \hat{\alpha} + \hat{\beta}_X X + \hat{\lambda}_2(1) + \hat{\lambda}_3(0) + \hat{\delta}_2 X(1) + \hat{\delta}_3 X(0)\\ &= \hat{\alpha} + \hat{\beta}_X X + \hat{\lambda}_2 + \hat{\delta}_2 X,\\ &= \left(\hat{\alpha} + \hat{\lambda}_2\right) + \left(\hat{\beta}_X + \hat{\delta}_2\right)X,\\ &= \hat{\alpha}' + \hat{\beta}' X.\end{aligned}$$

Similarly for the third category, $Z_3 = 1$:

$$\begin{aligned}\hat{Y} &= \hat{\alpha} + \hat{\beta}_X X + \hat{\lambda}_2(0) + \hat{\lambda}_3(1) + \hat{\delta}_2 X(0) + \hat{\delta}_3 X(1),\\ &= \hat{\alpha} + \hat{\beta}_X X + \hat{\lambda}_3 + \hat{\delta}_3 X,\\ &= \left(\hat{\alpha} + \hat{\lambda}_3\right) + \left(\hat{\beta}_X + \hat{\delta}_3\right)X,\\ &= \hat{\alpha}'' + \hat{\beta}'' X.\end{aligned}$$

We have three linear equations for predicting *Y* from *X*. But they have differing constants and regression parameters. These differences, of course, stem from the effects of going from one level of *Z* to another.

The real meaning and importance of interaction can perhaps be seen here: interaction means that the nature and strength of a linear *Y*-*X* relationship as measured by the regression parameters depend on the level of another variable.

Check out more Helpful Hints at **edge.sagepub.com/johnson8e**

in X on the mean of Y, when controlling for or holding other variables constant. Most software offers the option of calculating unstandardized or standardized coefficients.

Table 14-15 presents a comparison of the regression using standardized and unstandardized variables.

Columns 2 through 7 show alternatively raw scores and scaled values for the Gini, labor, and union in our abbreviated (minus Japan) dataset. The standardized or scaled variables are calculated with the raw data with the method introduced in chapter 13. For instance, to convert or transform Y to a standardized score, *y*, use the formula

$$y_i = \frac{(Y_i - \bar{Y})}{\hat{\sigma}_Y},$$

where $\bar{Y}$ is the mean of Y and $\hat{\sigma}_Y$ is its standard deviation. (In common usage, lowercase letters denote standardized variables.)

The bottom of table 14-15 demonstrates the properties of standardization. Note that the means are zero while the standard deviations are 1.0. In essence, we have converted the original measurement scales (e.g., percentages) to ones that are now (in a statistical sense) comparable. Hence, whereas a one-unit change on the original union density scale means a 1 percent increase or decrease, it now means a one-standard-deviation change. Similarly, labor protections, which are measured on a 1 to 6 scale in the original units, now are measured on a scale in which the basic unit is a standard deviation. (Examine table 14-15 to glean further insights.)

Regression analysis now simply entails using scaled variables instead of raw data. The results of standardized regression will be the same as those of unstandardized regression in these two respects:[14]

- Many measures of fit (e.g., r, R^2) will be the same.
- The results of tests of significance (e.g., t and F statistics, probabilities) will be the same.

But they differ in these two ways:

- There is no constant in the standardized equation.
- The numerical values of the standardized regression coefficients will not be the same, but they will have the same sign.

The differences can be seen in the two estimated models from the data in table 14-15 (note the absence of Japan):

Raw (unstandardized): $\bar{Y}_i = 42.10 - .14$ Union $- 2.48$ Labor, $R^2 = .73$, $F_{2,17} = 23.33$.

Scaled (standardized): $\bar{Y}_{I=} - .66$ Union $- .44$ Labor, $R^2 = .73$, $F_{2,17} = 23.33$.

If the results are basically the same, why bother? For one thing, one runs across results based on scaled variates all the time in scholarly literature, and it is advantageous to be familiar with them. In addition, many computer programs routinely report standardized regression coefficients (sometimes called *beta weights* or simply *betas*). More significant, perhaps, the comparability of regression coefficients calculated from standardized data supposedly allows one to assess the "relative" importance of explanatory variables. The coefficient for union concentration in the second equation, –.66, presumably implies it is a slightly better predictor of inequality than is labor protection, the coefficient of which is –.44. This alleged advantage might be reflected in statements such as the following: "All else being

[14] If you want to clarify expressions like these, simply replace the variable's symbols and codes with substantive names. Thus, for example $P(Y = 0)$ can in the present context be read literally as "the probability that 'contributed' equals 'did not contribute.'"

TABLE 14-15 Raw and Standardized Inequality Data

Country	Gini Index	Scaled Gini Index	Labor Protection	Scaled Labor Protection	Union Density	Scaled Union Density
Australia	35.2	0.732	1.38	–0.946	23.1	–0.594
Austria	29.1	–0.686	2.41	0.407	35.7	0.009
Belgium	33.0	0.221	1.02	–1.418	55.6	0.962
Canada	32.6	0.128	1.91	–0.250	28.2	–0.350
Denmark	24.7	–1.709	2.29	0.250	72.5	1.771
Finland	26.9	–1.198	3.00	1.182	74.8	1.881
France	32.7	0.151	2.63	0.696	8.2	–1.307
Germany	28.3	–0.872	2.97	1.143	23.2	–0.589
Greece	34.3	0.523	2.11	0.013	24.5	–0.527
Ireland	34.3	0.523	1.39	–0.932	36.3	0.038
Italy	36.0	0.919	2.58	0.630	34.0	–0.072
Luxembourg	30.8	–0.291	3.39	1.694	42.3	0.325
The Netherlands	30.9	–0.267	2.23	0.171	22.4	–0.628
New Zealand	36.2	0.965	1.16	–1.235	22.6	–0.618
Norway	25.8	–1.453	2.65	0.722	53.0	0.837
Spain	34.7	0.616	3.11	1.326	16.2	–0.924
Sweden	25.0	–1.639	2.06	–0.053	78.0	2.034
Switzerland	33.7	0.384	1.77	–0.433	17.8	–0.848
UK	36.0	0.919	1.09	–1.326	29.2	–0.302
USA	40.8	2.035	0.85	–1.642	12.6	–1.097
Means	32.05	0	2.1	0	35.51	0
Standard deviations	4.3	1	0.761	1	20.89	20.89

equal, a one-unit (one-standard-deviation) increase in unionization gives a .66 unit reduction in Gini scores, while an identical change in labor protection produces only a .44 decline. Since both are measured by the same metric (standard deviations), union density really is a more important explanation."

The seeming comparability of the standardized coefficients tempts some scholars into thinking that the explanatory power of, say, *X*, can be compared with that of another independent variable, say, *Z*. It would be easy to conclude, for example, that if b_{YX} is larger in absolute value than b_{YZ}, the former might be a more important or powerful predictor of *Y* than the latter. (Remember, we are talking about the standardized coefficients, which now presumably have the same measurement scale.) Yet you should be extremely careful about inferring significance from the numerical magnitudes of these coefficients. Such comparisons of the "strength of relationship" are possible only to the extent that *all the original independent variables have a common scale or unit of measurement.* The standardization process just changes the variables to standard deviation scales. It does not change or enhance their substantive interpretation. Also, standardization is affected by the variability in the sample, as can be seen by noting the presence of the standard deviations in the above formula. So if one independent variable exhibits quite of bit of variation while another has hardly any at all, it may be wrong to say the first is a more important explanation than the second, even if its standardized coefficient is larger.[15]

We reemphasize two points. First, transforming variables by standardization just changes their measurement scales. It does not alter their interrelationships. Therefore, tests of significance and measures of fit are the same for both sets of data. This is apparent from the two equal R^2 values (that is, $R^2 = .73$ in both instances). This will always be the case. And the regression constant drops out of the equation when standardized variables are used.

Measuring the Goodness of Fit, R^2

A lot of ink has been spilled over the best way to assess the adequacy of linear models. It is safe to say that no single number will tell us all we need to know. Consequently, political scientists have to reach deeply into their toolbox for devices that show different aspects of the model. Many of these are relatively advanced, however, so we will stick with the much used (and abused) multiple correlation coefficient, R^2, which of course is the explained (by regression) sum of squares divided by the total sum of squares.

R^2 varies from zero to 1. R^2 never decreases as independent variables are added. But just throwing more variables into a model usually will not add to the understanding

[15] For essentially the same reasons, you might not want to compare standardized regression coefficients based on samples from two different populations. See John Fox, *Applied Regression Analysis, Linear Models, and Related Methods* (Thousand Oaks, Calif.: SAGE, 1997), 105–8.

of variation in Y. Each independent variable added must be carefully considered. Moreover, the number of variables in a model *cannot* exceed the number of data points, and, if they are equal, the model will fit perfectly and R^2 will be 1.0.

Logistic Regression

Suppose we want to explain why people in the United States do or do not contribute money to political causes. Is it mostly a matter of public spiritedness or partisan passion? Or do donations depend mostly on economic well-being? As we have suggested many times before, such a study should start from a theory or at least a tentative idea of political participation. We might hypothesize, for example, that demographic factors such as education, age, and income are related to participation: older, well-heeled, college-graduate whites will donate more frequently and generously than lower-status individuals do. Alternatively, we could propose that partisanship will trump social and economic factors: strong partisans will be generous no matter what their financial or social situation is. To test this proposition, we could collect measures of these concepts from a survey or poll.

Table 14-16 shows ten cases selected randomly from the United States Citizenship, Involvement, Democracy study that we have previously used for examples. Besides indicators of education, income, age, and so forth, it asked respondents if they had donated to a political organization in the last year. Replies are coded 0 for "no" and 1 for "yes," thus creating a binary or dichotomous dependent variable. The questionnaire also contained material from which we constructed a four-category ordinal variable of partisan feelings: nonpartisan, weak, moderate, and strong.

One might wonder how we could use a method like multiple regression to analyze these data, since, strictly speaking, the dependent variable is not numeric or quantitative. (Earlier we saw that categorical independent variables like partisanship can be coded as dummy variables and entered into regression equations along with quantitative variables.) Indeed, a major problem for the social scientist is to explain variation in dependent variables of this type. Consider, for instance, figure 14-6, which shows the plot of donation ("no" or "yes") by respondents' age.

Incidentally, this figure and subsequent analyses are based on a sample of 150 cases drawn randomly from the complete data file. This sample of a sample is called a "training" dataset, and we use it to develop and test models. When one seems to fit, we can apply it to the larger, "verification" data that remain in the original sample. Using a relatively small N often simplifies one's analysis. For one thing, we do not have to plot more than 1,000 points, which usually leaves a blob of ink on the page. Moreover, even trivial relationships can be statistically significant when N is large.

TABLE 14-16 Citizen Involvement in Democracy Data

Dependent Variable (*Y*)	Independent Variables (*X, W, Z*, . . .)			
Donated	**Income**	**Age**	**Partisanship**	**Education**
1	5	60	Strong	Post–high school
0	7	35	Nonpartisan	High school
0	4	55	Strong	Post–high school
0	7	56	Strong	College or more
0	9	57	Moderate	High school
0	8	44	Nonpartisan	High school
0	3	31	Moderate	High school
1	6	64	Strong	Post–high school
1	3	35	Strong	High school
0	9	59	Strong	Post–high school

Education, originally recorded with eight categories, has been recoded into one with four levels.

Source: Marc Morjé Howard, James L. Gibson, and Dietlind Stolle, "The U.S. Citizenship, Involvement, Democracy Survey," Center for Democracy and Civil Society (CDACS), Georgetown University, 2005.

In any case, we observe two parallel lines of dots that do not tell us much, if anything, about the relationship between contributing and age. One thing we can infer is that there are fewer "yes" than "no" responses.

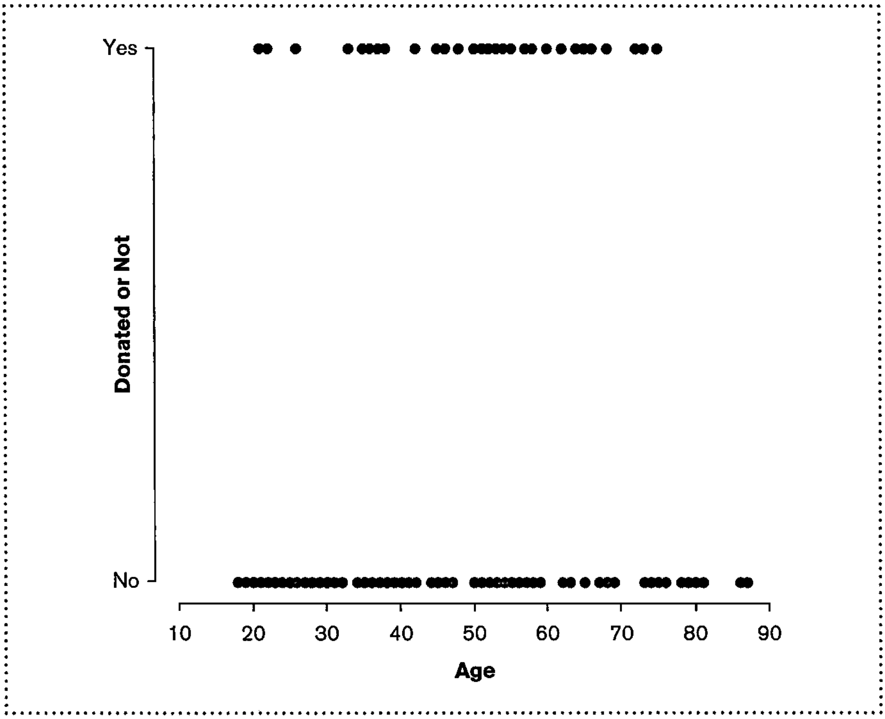
Table 14-17 shows the marginal distribution of this variable.

What to do? We might conceptualize the problem this way. Denote the two outcomes of the dependent variable, *Y*, as 1 for "yes" and 0 for "no." Each person in the study, in other words, is assigned a score on the dependent variable of 1 or 0, depending on whether or not that person contributed. For a number of reasons, this type of response variable creates problems for ordinary regression analysis. As a consequence, we often do not analyze *Y*, per se, but rather some function of it. That is, the dependent variable is not *Y* with its two values but *Y*′, which is a function of *Y*.

TABLE 14-17 Marginal Distribution of Political Contributions

Response	Frequency	Proportion
No	116	.77
Yes	34	.23
Total	150	1.0

FIGURE 14-6 **Contributed to Political Organizations by Age**



Source: US Citizenship, Involvement, Democracy Survey.

When confronted with binary responses such as "no" and "yes," we can slightly reconceptualize the situation as one of predicting a "no" or "yes" answer. To do so, interpret the expected value of *Y* as "the probability that *Y* equals 1" because

$$E(Y) = [1 \times P(Y = 1)] + [0 \times P(Y = 0)] = P(Y = 1).$$

Note that *P*(*Y* = 1) means "the probability that *Y* equals 1," which in this context is the probability that a person donated. Similarly, *P*(*Y* = 0) is defined as the probability of not giving.[16] (Frequently, the generic terms *success* and *failure* are employed to describe these probabilities, as in "The probability of success is *P*, and the probability of failure is 1 – *P* = *Q*.) As noted before, the expected value of a variable can be thought of roughly as the sum of its possible values times the probabilities of their occurrence.[17]

[16] If you want to clarify expressions like these, simply replace the variable's symbols and codes with substantive names. Thus, for example, *P*(*Y* = 0) can in the present context be read literally as "the probability that the variable 'donated' equals 'did not contribute.'"

[17] More precisely, the expected value of a probability distribution is called the mean of the distribution.

Therefore, our job is to understand and predict probabilities, not raw scores as in the inequality models.

So what can be done? We could simply treat Y as a numerical or quantitative dependent variable and use it in the normal regression analysis described above. Such a procedure is called a linear probability model, and this model "looks" just any other regression equation:

$$E(Y) = P = \hat{\beta}_0 + \hat{\beta}_1 X.$$

The linear probability model works reasonably well when all predicted values lie between .2 and .8, but statisticians still believe that it should not generally be used. One reason is that the predicted probabilities can have strange values, since the linear part of the model can assume just about any value from minus to plus infinity, but a probability by definition must lie between zero and 1. So, for example, the estimated probability of voting might be 1.3. The estimate might be valid but has no meaning. (How can a probability of something happening be greater than 1?) In addition, the linear probability model violates certain assumptions that are necessary for valid tests of hypotheses. The variance of the error term (ε) in the model, for example, violates the assumptions of homoscedasticity and normal errors, and the results of a test of the hypothesis that a β is zero might be suspect. For these and other reasons, social scientists generally do not use a linear probability model to analyze dichotomous dependent variables.

We certainly do not want to give up, because many dichotomies or binary dependent variables or responses are frequently worth investigating. A common solution is to use **logistic regression** analysis that at first blush appears to either have a strange dependent variable *or* an even stranger equation. (You can easily move from one form to another.) The apparent "weirdness" arises because we can either use a nonlinear equation to explain Y (or probability of success) or a linear model to explain a function of Y, so to speak. Table 14-18 lays out the choices. (We explain odds and log odds a little later.)

The logistic regression function for two independent variables, X_1 and X_2, and a dichotomous dependent variable, Y, has the form

$$Prob(Y = 1) = \hat{P} = \frac{e^{(\hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2)}}{1 + e^{(\hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2)}}.$$

This rather mysterious-looking formula can be easily understood simply by looking at some graphs and making a few calculations. First note that e, which is often written *exp*, stands for the exponentiation function. A function can be thought of as a machine: put a number in, and another, usually different number comes out. In

TABLE 14-18 **Modeling Binary Responses**

Type of Regression	Form of the Dependent Variable	Form of Model	Model
Linear probability	Probability $Y = 1$	Linear	$E(Y) = P = \hat{\beta}_0 + \hat{\beta}_1 X.$
Logistic	Probability $Y = 1$	Nonlinear	$Prob(Y = 1) = \hat{P} = \frac{e^{(\hat{\beta}_0 + \hat{\beta}_1 X)}}{1 + e^{(\hat{\beta}_0 + \hat{\beta}_1 X)}}.$
	Odds: $Y_i' = \frac{p}{(1-p)}$	Nonlinear	$Y_i' = e^{(\hat{\beta}_0 + \hat{\beta}_1 X)}$,
	Log odds (logit): $Y_i'' = \ln\left(\frac{p}{(1-p)}\right)$	Linear	$Y_i' = \hat{\beta}_0 + \hat{\beta}_1 X$

ln = natural logarithm, e = exp = exponential function.

this case, since *e* is a number that equals approximately 2.718218, *X* enters as the exponent of *e* and emerges as another number, 2.71828^X. For instance, if *X* equals 1, then e^1 is (approximately) 2.7182, and if $X = 2$, e^2 is about 7.3891. (Many hand-held calculators have an exponentiation key, usually labeled e^X or *exp*[*X*]. To use it, just enter a number and press the key.) Although this function may seem abstract, it appears frequently in statistics and mathematics and is well known as the inverse function of the natural logarithm; that is, $\log(e^x) = x$. For our purposes it has many useful properties.

Ways of Thinking about Dichotomous Dependent Variables

In bivariate and multiple regression analysis the dependent variable (*Y*) is quantitative or numerical, and one statistical goal is to explain its variation. But when *Y* has just two categories (such as 1 and 0) there are a couple of ways of setting up and interpreting models. One approach is to examine *Y* directly by modeling the probability that *Y* equals 1 or zero. These models have regression-like coefficients for the *X*s, but they appear in the exponents of somewhat complicated-looking equations for the probabilities and cannot be understood in the simple "a-one-unit-change-in-*X*-produces-a . . ." framework of ordinary regression. So understanding the meaning of logistic coefficients is not intuitive.

It is possible, however, to model, not the probability that *Y* equals 1, but the *odds* that *Y* equals 1 as opposed to zero. (The odds that *Y* equals 1 are not the same as the probability that Y equals 1, as we emphasize later in the chapter.) In this formulation, the odds become a kind of dependent variable, and the analyst's objective is to study what affects it. Furthermore, it is frequently convenient to transform the odds by taking their natural logarithm to get "logits." So logits, too, can be considered as a sort of dependent variable. The use of logits is popular because models for them are linear in the explanatory factors, and a (partial) logistic regression coefficient does have the interpretation that a one-unit change in X is associated with (partial) beta-unit change in the logit or log odds when other Xs have been controlled. The difficulty, of course, is that now the meaning of the dependent variable—a logit—is not obvious. Fortunately, all these formulations are equivalent, and it is possible to move back and forth among them. The first part of this logistic regression section develops and explains models for probabilities, and a latter part looks at models for the log odds.

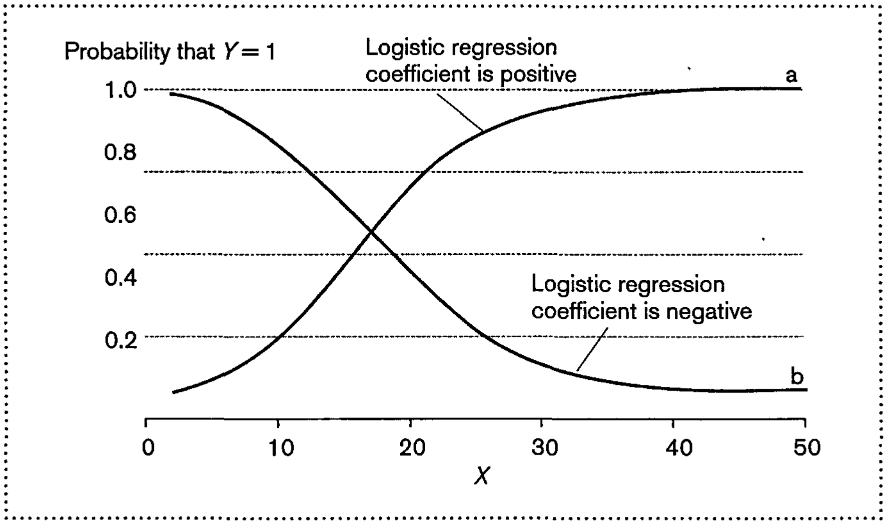
The logistic function can be interpreted as follows: the probability that Y equals 1 is a nonlinear function of X, as shown in figure 14-7. Curve a shows that as X increases, the probability that Y equals 1 (the probability that a person votes, say) increases. But the amount or rate of the increase is not constant across the different values of X. At the lower end of the scale, a one-unit change in X leads to only a small increase in the probability. For X values near the middle, however, the probability goes up quite sharply. Then, after a while, changes in X again seem to have less and less effect on the probability, since a one-unit change is associated with just small increases.

Depending on the substantive context, this interpretation might make a great deal of sense. Suppose, for instance, that X measures family income and Y is a dichotomous variable that represents ownership or nonownership of a beach house. (That is, $Y = 1$ if a person owns a beach house and 0 otherwise.) Then for people who are already rich (that is, have high incomes), the probability of ownership would not be expected to change much, even if they increased their income considerably. Similarly, people at the lower end of the scale are not likely to buy a vacation cottage even if their income rises substantially. It is only when someone reaches a threshold that a one-unit change might lead to a large change in the probability.

Curve b in figure 14-7 can be interpreted the same way. As X increases, the probability that Y equals 1 decreases, but the amount of decrease depends on the magnitude of the independent variable.

The essence of nonlinear models is that the effects of independent variables are not constant but depend on their specific values. So the logistic regression function has a reasonable interpretation. It also meets the objectives mentioned earlier—namely, that predicted values will lie between 0 and 1, which are the minimums

FIGURE 14-7 **Logistic Functions**



Note: Hypothetical data.

and maximums for probabilities, and that the assumptions of hypothesis testing will be met.[18]

Logistic regression can be further understood with a numerical example. Using a procedure to be described shortly, the estimated logistic regression equation for the participation and democracy survey data with donated ("no" or "yes") as the dependent variable and age and income as predictors is

$$\hat{P} = \frac{e^{-2.89+.02\,\text{Age}+.18\,\text{Income}}}{1+e^{-2.89+.02\,\text{Age}+.18\,\text{Income}}}.$$

In this particular equation, $\hat{\beta}_0$ equals –2.89, β_1 equals .02, and β_2 equals .18. These numbers are called **logistic regression coefficients**, which are related to multiple **regression coefficients** in that they show how the probability of voting changes with changes in the independent variable.

Although an explanation of how the coefficients were calculated goes beyond the scope of this book (and computer programs for doing the work are widely

[18] Of course, like any statistical technique, logistic regression analysis assumes certain conditions are true and will not lead to valid inferences if these conditions are not met.

available), we can start to examine their meaning by substituting some values for the independent variables into the equation. Keep in mind, however, that logistic regression coefficients (the βs) are similar to regular regression coefficients: they indicate the effect that a change in a particular independent variable produces when the other independent factors in the model have been held constant. They are like partial coefficients of multiple regression because each isolates the impact of a specific *X* net of all the other *X*s in the equation. But remember that a β does not have a simple linear effect on *Y* for a given change in *X*. It instead is nonlinear in its consequences. This interpretation becomes clearer as we go on.

Consider a person who reports zero income and age ($X_1 = X_2 = 0$). Then the equation becomes

$$\begin{aligned}\hat{P} &= \frac{e^{-2.89+.02(0)+.18(0)}}{1+e^{-2.89+.02(0)+.18(0)}}\\ &= \frac{e^{-2.89}}{(1+e^{-2\ 89})}\\ &= .05.\end{aligned}$$

This expression means that the estimated probability that a person zero years old and without any income will donate to a political cause is .05, signifying little or no chance at all. (This probability perhaps makes sense for someone who is not born and has no income; its main value, however, is as a baseline that can be compared with the results for a 70-year-old person (X_1=70) in the highest income category (X_2 = 11). (We are using the category labels as an actual interval variable.) The predicted probability of donating is now

$$\begin{aligned}\hat{P} &= \frac{e^{(-2.89+.02(70)+.18(11))}}{1+e^{(-2\ 89+.02(70)+.18(11))}}\\ &= \frac{e^{(.49)}}{1+e^{(.49)}}\\ &= .62.\end{aligned}$$

Here we see that an individual with these characteristics has a better than 60 percent chance of contributing. Similar substitutions show how different combinations of the independent variables change the probability. Let's fix (hold constant) income at its mean value (5.09) and let age vary from 20 to 90 by 10-year intervals. We can stick those numbers one after another into the estimated model to produce a table of predicted values (see table 14-19).

Using the table, we can see that for a fixed value of income, the probability of a contribution increases with age. (Always pay attention to the sign of the coefficients.) Yes, the table tells us that age is positively related to donating, but what about income? Its coefficient also has a positive sign, so we assume that probabilities will increase as income goes up, for fixed values of age. Another simple table reveals that this is so (see table 14-20).

TABLE 14-19 **Effects of Varying Age While Holding Income Constant on Predicted Donations**

Income	Age	Predicted Probability of Donating
5.09	20	0.17
5.09	30	0.20
5.09	40	0.24
5.09	50	0.27
5.09	60	0.32
5.09	70	0.36
5.09	80	0.41
5.09	90	0.46

With age set at its median value (45 years), we see that as income increases, so too does the predicted probability. Both results make sense. Older and wealthier people are generally more civic minded than, say, the poor and young, so we would expect them to be more likely to donate. What about the effects of both variables simultaneously? Can they be visualized? Figure 14-8 shows a particularly simple but instructive way to graph these. First look at the *x*-axis, which is marked off by ages; the *Y* scale is just the predicted probability of making a political donation. The points (symbols) stand for the predicted probability for cases with specific category combinations of the independent variables.

To appreciate what the lines and dots mean, consider the following:

- Since the *x*-axis is age, the figures suggest that as age increases, so too does the predicted probability.
- This pattern is true for *all* three levels of income shown (minimum, mean, and maximum).
- For any particular age, the higher one's income, the higher the probability of donating.
- The difference in effects of "income" on probability between adjacent income levels is more or less constant across income. (There is no apparent interaction effect.)

To take a quick example, look at the left side of the graph, where predicted probabilities for the youngest (age = 18) group lie. Notice that as you jump up from one income level to the next, the probability increases. This means income is positively related to the chances of making a political contribution (remember the positive beta?). And this effect is not spurious because of age, for we have held age constant: at each age, the previously described relationship holds. Now, select an income level, say the mean (middle line). Notice that the line slopes slightly upward, indicating a

TABLE 14-20 **Effects of Varying Income While Holding Age Constant on Predicted Donations**

Income	Age	Predicted Probability of Donating
1	45	0.14
2	45	0.16
3	45	0.19
4	45	0.22
5	45	0.25
6	45	0.29
7	45	0.33
8	45	0.37
9	45	0.41
10	45	0.45
11	45	0.50

positive relationship: as age increases, the probability of making a contribution also increases slightly. The same is true for the other two income groups. Conclusion: age, too, positively affects the propensity to give. Finally, take note of the fact that the lines are more or less parallel. In words, this means that the effect of age on donating is the *same* (direction and strength) at all levels of income. Conversely, the income-probability connection shows little or no interaction.

If we wanted to check for possible interaction effects, we could simply add an interaction variable, Z = age × income, estimate its coefficient, and test to see if it differs from zero. We do some of this in a moment.

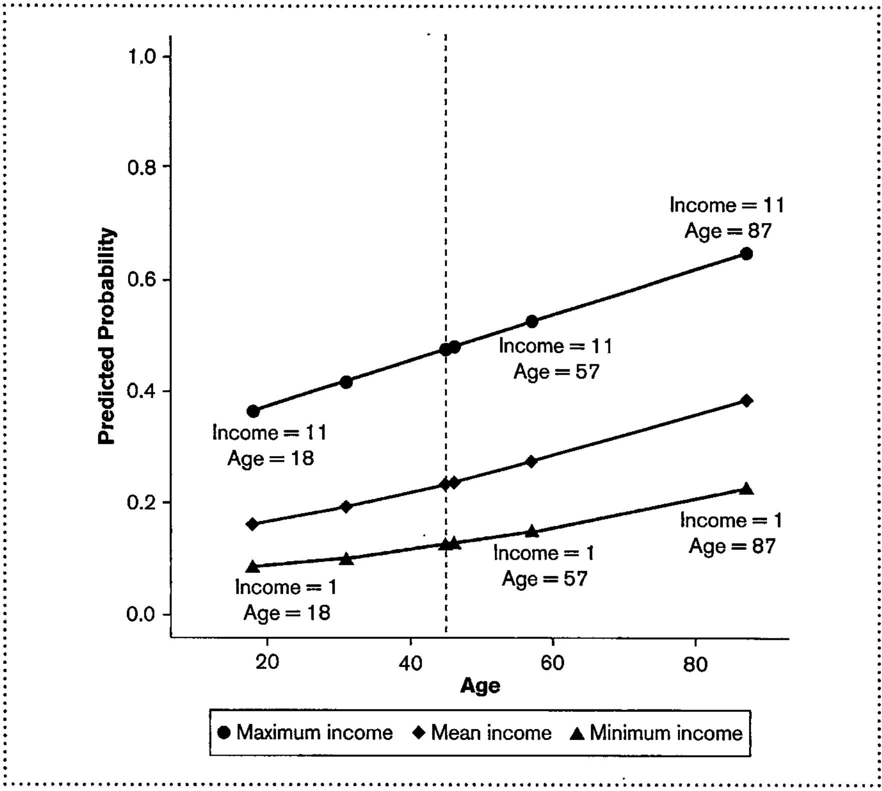
Estimating the Model's Coefficients

It is natural to wonder how the coefficient estimates are derived, and it would certainly simplify things if we could provide straightforward formulas for calculating them. Unfortunately, there are no such easy equations. Instead, logistic regression analysis is best performed with special computer programs. Logistic regression has become so widely used that the appropriate tools can be found in many statistical program packages such as SPSS, MINITAB, R, Stata, and SAS. Your instructor or computer consultant can help you find and use these programs. We recommend that if you have a dependent variable with two categories and want to perform regression, ask for a logistic regression program.[19]

Although the details are beyond the scope of the book, the method used to estimate unknown coefficients relies on a simple idea: pick those estimates that maximize the likelihood of observing the data that have in fact been observed. In effect, we propose a model that contains certain independent variables and hence unknown coefficients. Associated with the model is a *likelihood function, L*. The parameters in the function L give the probability of the observed data. That is, the data points are treated as fixed or constant, and the likelihood is a function of unknown parameters. Using the principles of differential calculus, a numerical algorithm selects values of the parameters that maximize L. Logically enough, they are called

[19] Quite a few methods can be used to analyze these kinds of data. A related procedure, called probit analysis, is widely used, and if the data are all categorical, log-linear analysis is available.

FIGURE 14-8 **Effects of Independent Variables on Predicted Probabilities**



maximum-likelihood estimators. Therefore, the aim of the behind-the-scenes number crunching is to find those values for the parameters that maximize the probability of obtaining the observed data that we did.

For computational purposes, the logarithm of the likelihood function is calculated to give the *log likelihood function*, or *LL*. Keep an eye open for it because log likelihood functions, which are somewhat analogous to sums of squares in regression, appear in many model-fitting and testing procedures, as we will see in the next section.

If the estimated coefficients are calculated correctly and certain assumptions are met, they have desirable statistical properties. They are, for instance, unbiased estimators of corresponding population parameters and can be tested for statistical significance.

Measures of Fit

As in the case of simple and multiple regression, researchers want to know how well a proposed model fits the data. The same is true of logistic regression. After estimating a model, we want to know how well it describes or fits the observed data. Unfortunately, there is no universally accepted summary measure like R^2 that describes in an intuitively appealing way the agreement of a model and data. Several measures of goodness of fit exist, although they have to be interpreted cautiously. In fact, considerable disagreement exists about which measure is best, and none of the alternatives has the seemingly straightforward interpretation that the multiple regression coefficient, R^2, has.

With ordinary regression analysis, one way to calculate the fit is to compare predicted and observed values of the dependent variable and measure the number or magnitude of the errors. Alternatively (and equivalently), we can determine what proportion of the variation in Y is statistically explained by the independent variables. In the case of binary dependent variables, there are no precise analogs to total and residual sums of squares.

Logistic regression involves roughly analogous steps, but the procedures are a bit more complicated and cumbersome, so we simply sketch out the general ideas. Our main objective is to provide a working understanding of substantive research articles and computer output.

Most logistic regression software programs routinely report the values of log likelihood functiôns, *LL*. (They will be negative numbers.) Occasionally, as with the popular program package SPSS, the result given is –2 times the log likelihood, but you can switch back and forth easily by the appropriate multiplication or division. As an example, the log likelihood for the logistic regression of age and income on probability of donating is $LL = -83.506$. This number looks large, but what exactly does it mean? Unfortunately, the number is not terribly informative by itself. But it can be compared with the *LLs* obtained for other models. And these comparisons can be used to gauge the overall fit and test hypotheses about sets of coefficients.

A simple strategy for assessing fit is to contrast the log likelihood of a model with one having only a constant term, LL_0, with a model that contains, say, two independent variables, X_1 and X_2. This log likelihood we denote LL_C, for "current" model. A measure of "improved" fit, the so-called pseudo-R^2, compares the log likelihood from the null model (only an intercept) to the log likelihood from the full model (all covariates included), then, is

$$R^2_{\text{pseudo}} = \frac{LL_0 - LL_C}{LL_0},$$

where LL_0 is the log likelihood for the null or "reduced" model and LL_C is the complete or "full" model. The denominator plays the role of the total sum of squares, while the numerator shows the difference in the fit when independent variables have been added and might be loosely considered the "explained" portion. The "pseudo" in the resulting R-squared indicates that this statistic is not the same as the R^2 of ordinary regression, and it certainly does not represent explained variation. But the basic idea is the same: pseudo-R^2 roughly suggests the relevance of a set of independent variables in understanding the probability that $Y = 1$. Moreover, we make use of log likelihoods (LL) in a moment.[20] For the citizenship example, LL_0 for the model with no independent variables (only a constant term) is –87.60, and LL_C for the model with age and income included is –83.51. Thus, the pseudo-R^2 is

$$R^2_{\text{pseudo}} = \frac{\left[(-87.60)-(-83.51)\right]}{-87.60} = \frac{-4.10}{-87.60} = .05.$$

It can be easily calculated because the log likelihoods are routinely reported. This number suggests that the addition of two independent variables did not improve the fit very much. But before rejecting the model, keep in mind that the pseudo-R^2 is not an infallible indicator of fit and that others have been proposed.[21] More important, as we have said, many statisticians are wary of giving any "explained variation" interpretation to logistic regression results and don't bother with pseudo-R^2 or its many variants. Perhaps because logistic regression has been incorporated into standard political analysis relatively recently, there is no widely accepted and used list of measures. Some authors provide several indicators, whereas others give few. Thus, when reading articles and papers that use dichotomous dependent variables and logistic regression, you may have to reserve judgment about the evidence in favor of a particular model.[22]

Significance Tests

We return to contrasting nested models as a way to assess both overall models and individual coefficients. But the residual or error sum of squares is replaced by *deviance,* which is usually defined as "minus twice the log likelihood": $D = -2LL$. Think

[20] A good review and proposal for such a measure is Tue Tjur, "Coefficients of Determination in Logistic Regression Models—A New Proposal: The Coefficient of Discrimination," *American Statistician* 63, no. 4 (2009): 366–72.

[21] See, for example, J. Scott Long, *Regression Models for Categorical and Limited Dependent Variables* (Thousand Oaks, Calif.: Sage, 1997), 104–13. Note also that some statisticians recommend against using most R^2-type measures in logistic regression work. See, for example, David W. Hosemer and Stanley Lemeshow, *Applied Logistic Regression Analysis* (New York: Wiley, 1989), 148.

[22] The same comment applies to any data analysis technique: empirical results have to be interpreted and accepted with caution.

of deviance as being analogous to the residual sum of squares. The objective is to find a combination of explanatory variables that make it as small as possible. One way to see if we are making progress is to compare the deviance of models with and without a particular variable, *X*. If the difference is "large" as judged by some standard, the term significantly improves the fit and is retained; otherwise, it is dropped. The *LL* and deviance too frequently tumble out of logistic regression software, so we do not need to dwell on its computation. Instead, we just use the deviances from various models to evaluate the significance of coefficients. What results resembles, but is not equivalent to, ANOVA tables.

TESTING THE OVERALL MODEL. To start, we can perform a test analogous to the *F*-test in multiple regression.[23] This procedure follows the steps in the previous section. Let LL_C be the log likelihood for a current or "complete" model—the one with all the explanatory variables of interest included—and let LL_0 be the log likelihood for the "reduced" model—the one with one or more independent variables eliminated or, as the case arises, the model with no explanatory factors, only a constant. Then the difference between the two likelihoods forms a basis for a test of a test statistic, the likelihood ratio chi square (*LRX*) with degrees of freedom equal to the difference in the number of parameters in the model:

$$LRX = -2(LL_0 - LL_C).$$

LRX can be recast as the difference in deviances, since $D = -2LL$:

$$LRX = D_0 - D_C,$$

where D_0 and D_C are the deviances for the reduced and full models, respectively. For large samples and under the modeling assumptions, *LRX* has a chi-square distribution with $df = K - p$, where *K* is the number of independent variables in the full model and *p* is the number in the reduced equation. This observed statistic tests the null hypothesis that a β or a set of βs is zero. It can be used to test one coefficient at a time, in which case the number of degrees of freedom is 1. A small *LRX* (that is, near zero) means the "tested" coefficients are not statistically significant and perhaps should not be included, whereas a large one suggests that they may be (statistically) important.

Let's apply the likelihood ratio chi-square test to the problem we have been working on, understanding why people contribute to political causes. So far, our model only contains demographic data (income and age), but we can expand it later. First let's

[23] Consider, for example, a model that contains two types of variables—one group measuring demographic factors and another measuring attitudes and beliefs. The investigator might want to know if the demographic variables can be dropped without significant loss of information.

do a global test. The software program we use throughout the book, R, gives us the results in table 14-21.

Here we are testing the null hypothesis that $\beta_{Age} = \beta_{Income} = 0$ against the alternative that at least one population coefficient is not zero. The test statistic turns out to be *LRX* = 8.19, with 2 degrees of freedom. (The degrees of freedom for testing nested models is just the difference in degrees of freedom for the models or, what is the same, the number of variables in the full model minus the number in the reduced model. Here it is: 2 – 0 = 2 degrees of freedom.) The critical values for a chi-square statistic at the .05 and .01 levels are 5.99 and 9.21, respectively. The observed *LRX* lies between them (i.e., $5.99 < LRX = 8.19 < 9.21$), so we know it is significant at the .05 level but not at .01. The table shows that the attained probability is actually .02. All this testing means simply that *if* the null hypothesis of no effects of age and income is true, the probability of the observed result (8.19) (or one even larger) is about 2 in 100.

The overall model with both coefficients is significant, but we don't know if just one or both coefficients contribute to the effect. For that we need a test of the individual parameters (see table 14-22).

The interpretation of the table follows. We are now interested in knowing if a particular partial regression coefficient is zero. First calculate *LRX* from either the log likelihoods or the deviances. (The choice depends only on what your software provides.) As in the other hypotheses tests, if the observed statistic exceeds the critical chi square at a specified level of significance, we reject the null hypothesis (that is, H_0: $\beta_j = 0$) in favor of the alternative ($\beta_j \neq 0$). The first test (*LRX* = 2.09) is not significant. Normally, we would stop here because including age does not help predict behavior. But simply to keep the example going, we next added income to see

TABLE 14-21 Likelihood Ratio Chi-Square Test

Model	Parameters in Model	Log Likelihood of Model (*LL*)	Deviance (–2 × *LL*)	$LRX = D_0 - D_C$	$df = K - p$	Prob.
Null or intercept only	$\hat{\beta}_0 = -1.03$	–87.60	–175.21	–	–	–
Complete	$\hat{\beta}_0 = -2.89$ $\hat{\beta}_{Age} = -.02;$ $\hat{\beta}_{Income} = .18$	–83.51	–167.01	8.19	2	0.02

$K = 2$, $p = 0$.

TABLE 14-22 **Likelihood Ratio Chi-Square Test of Individual Parameters in the Model**

Model	Parameters in Model	Log Likelihood of Model (*LL*)	Deviance (−2 × *LL*)	$LRX = D_0 - D_c$	$df = K - p$	Prob.
Null or intercept only	$\hat{\beta}_0 = -1.03.$	−87.60	−175.21	–	–	–
Age	$\hat{\beta}_0 = -1.81;$ $\hat{\beta}_{Age} = .02.$	−86.56	−173.12	2.09	1	.15
Age + Income	$\hat{\beta}_0 = -2.89;$ $\hat{\beta}_{Age} = .02;$ $\hat{\beta}_{Income} = .1.8.$	−83.51	−167.01	8.2	1	0.01

df_0, df_c = degrees of freedom for the reduced and complete models.

$K = 2$, $p = 0$.

Note: These results were calculated with R on weighted data

if it improves the model. We see that the second *LRX* (6.11) has a low probability (under the null hypothesis that $\beta_{Income} = 0$), and we conclude that income has a statistically significant impact on willingness to donate.

WALD TESTS. Articles in the scholarly literature frequently report significance tests for the individual coefficients using a different statistic. In the case of logistic regression, we usually want to test the hypothesis that in the population, a β equals zero. As an example, we might want to test the null proposition that the partial logistic coefficient relating education to income is zero. The form of this kind of test is roughly similar to the others we have described throughout the book: divide an estimated coefficient by its standard error. In this case, if the sample size is large (say, greater than 200), the result gives a statistic, *z*, which when squared has a chi-square distribution with 1 degree of freedom. That is,

$$z = \left(\frac{\hat{\beta}}{\hat{\sigma}_{\hat{\beta}}} \right),$$

where $\hat{\beta}$ is the estimated coefficient and $\hat{\sigma}_{\hat{\beta}}$ is its estimated standard error. When squared, this quotient, often labeled a "Wald" statistic, can be compared with a chi-square statistic with 1 degree of freedom. The test follows the usual path: establish a critical value under the null hypothesis that a β equals some value, compare the observed z to the critical value, and make a decision. (Recall that critical values for chi square can be found in appendix A.)

Software invariably reports the coefficients and their standard errors and usually the z or Wald statistic as well, so we need not worry about computing them by hand.

We conclude this section by pointing out that the accuracy of the Wald (or z) statistic depends on many factors, such as the sample size. As a result, some statisticians advise using the *LRX* statistic applied to one coefficient at a time. That is, test a model with K independent variables (and, hence, K coefficients) against one with $K - 1$ parameters. (The former would be the "current" model, the latter the "reduced" model.) If the difference is significant, the variable left out should perhaps be included. Otherwise, we might not reject the hypothesis that its coefficient is zero. But since the z or z^2 appears so frequently, it is important to be aware of its purpose.

An Alternative Interpretation of Logistic Regression Coefficients

We might summarize this point by saying that logistic regression analysis involves developing and estimating models so that the probability that Y equals 1 (or 0) is a *nonlinear* function of the independent variable(s):

$$P(Y = 1) = \text{Nonlinear function of } X.$$

It is possible, though, to rewrite the logistic regression equation to create a linear relationship between the Xs and Y. Doing so provides an alternative way to interpret logistic regression results. Instead of explaining variation in Y with a linear probability model or P with a logistic regression, we can work with odds, which are the probability of one response or value of a variable over the probability of another response or value of a variable, and use them as a dependent variable.

Suppose we sampled a person at random from a group of Americans. We could ask, "What is the probability (P) that this individual made a contribution to a political group or organization in the past year?" or, a related question, "What are the *odds* that this individual gave?" Probability and odds are not the same, for the odds are the ratio of *two* probabilities, the probability of donating compared with the probability of not donating:

$$\text{Odds} = O = \frac{P_{\text{Donate}}}{\left(1 - P_{\text{Donate}}\right)},$$

where P_{Donate} is the probability of voting.

Some examples will help to illustrate the difference. Suppose the probability that a randomly selected citizen makes a contribution is .2. Then the *odds* of her doing so are .2/(1 − .2) = .2/.8 = .25, or, as is commonly said, .25 to 1 or, more commonly, 1 out of 4. The person, in other words, is four times as likely *not* to donate as to make a contribution. As another example, suppose the probability of, say, voting is .8; then the *odds* are .8/(1 − .8) = .8/.2 = 4, or about 4 to 1. In this case, the citizen is more likely to vote than not to vote. In both examples, the terms in the denominator of the fraction are just 1 − *P*, which is the probability of not voting. (Since probabilities must add to 1—either a person did or did not vote—the probability of not voting is 1 − *P*.)[24] It is important not to confuse probabilities and odds; they are related, but not the same.

More generally, consider a variable, *Y*, that takes just two possible values, 0 and 1. Let *P* be the probability that $Y = 1$ and $Q = 1 - P$ be the probability that $Y = 0$. Then the odds that *Y* is 1 as opposed to 0 are

$$O = \frac{P}{(1-P)} = \frac{P}{Q}.$$

The term *O* has intuitive appeal, since it accords with common parlance. The odds, *O*, can vary from zero to infinity. If $O = 1$, then the "chances" that $Y = 1$ or 0 are the same—namely, 1 to 1. If *O* is greater than 1, the probability that $Y = 1$ is greater than 1/2, and conversely if *O* is less than 1, the probability is less than 1/2. Table 14-23 shows a few more examples of probabilities and odds in a case in which a random process can lead to just one of two possible outcomes.

TABLE 14-23 **Probabilities and Odds**

Probabilities	Odds
1.0	∞
.7	2.333
.5	1
.4	0.667
.1	0.111
0	0

Note: Read the odds as "*X* to 1."

Why bother with odds? Take a look at the logistic model. It is really a formula that relates *P* to some *X*s, so we ought to be able to rewrite it by putting 1 in front to obtain 1 − *P*. Then we could put the two equations together to get an expression for *P* over 1 − *P*. Here is how. To simplify, let $Z = \beta_0 + \beta_1 X_1 + \beta_2 X_2$. Now an expression for *P* can be written

$$P = \frac{e^Z}{1 + e^Z}.$$

[24] That is, $P + (1 - P) = 1$.

In the same fashion, we can write 1 – P as

$$1-P=1-\frac{e^{z}}{1+e^{z}}.$$

This latter expression can be simplified to

$$1-P=\frac{1}{1+e^{z}}.$$

Now we can put the two equations for P and 1 – P together to obtain an expression for the odds, $O = P/(1 - P)$:

$$O=\frac{P}{(1-P)}=\frac{\frac{e^{z}}{1+e^{z}}}{\frac{1}{1+e^{z}}}.$$

This expression in turn simplifies to

$$O = e^{z}.$$

Remember that we let $Z = \beta_0 + \beta_1 X_1 + \beta_2 X_2$, so this expression is really

$$O = e^{\left(\beta_0 + \beta_1 X_1 + \beta_2 X_2\right)}.$$

We have thus found a simple expression for the odds. It is still nonlinear because of the exponentiation, e. But a property of the exponentiation function is that $\log({}_eZ) = Z$, where log means the natural logarithm. So we find that the logarithm of the odds—called the log odds, or logit—can be written as a linear function of the explanatory variables:

$$Logit = \log O = \beta_0 + \beta_1 X_1 + \beta_2 X_2.$$

In essence, we have three versions of the dependent variable:

a. $Y = P$, a *probability* (nonlinear model)

b. $Y = O$, an *odds* of "success" (multiplicative model)

c. $Y =$ log-O, a *logit* or log odds (additive model)

Back to the logit for a moment. It can be interpreted in the same terms as multiple linear regression *if we keep in mind that the dependent variable is the logit, or log odds, not* Y *or probabilities.*

Refer, for instance, to our two-variable model of political contributions. We see that it can be written three ways, all of which lead to the same statistical conclusions. The choice depends on how comfortable one is with interpreting particular forms of the dependent variable.

a. $\hat{Y}_i = \frac{e^{-2.89+.02\text{Age}+.18\text{Income}}}{1+e^{-2.89+.02\text{Age}+.18\text{Income}}}$

b. $\hat{Y}_i = \hat{O} = e^{-2.89+.02\text{Age}+.18\text{Income}}$

c. $\hat{Y}_i = \text{Log-}\hat{O} = -2.89 + .02\text{Age} + .18\text{Income}$

Again, the easiest way to get a handle on what one of these equations is telling you is to make some substitutions. We could, for example, alter our previous perspective by switching from probabilities to odds. What are the odds of individuals with certain traits making monetary contributions to politics? How are they affected by changes in these characteristics? We look to equation (b). It can be reexpressed as

$$\hat{Y}_i = \hat{O} = e^{-2.89+.02\text{Age}+.19\text{Income}} = e^{-2.89}e^{.02}e^{.18}.$$

This means that the effects of changes in *X* are *multiplicative,* not additive. A one-unit jump in income produces an exponential change in the odds, *exp*(.18) = 1.20. That is, as income goes up one unit, the odds of making a political contribution *multiply* by 1.20. To get a handle on what this means, it is easiest to compute the estimated odds for several combinations of the independent variable, just as we did before. (As a matter of fact, we could just transform the estimated probabilities in the previous tables.) Suppose we hold age constant at 45 years and let income vary through its range. (Remember, we are treating income as a quantitative variable even though it is in fact an ordinal categorical variable with eleven ordered categories.) Start with someone in the first income group:

$$\hat{O} = e^{-2.89}e^{.02(45)}e^{.18(1)} = e^{-1.81} = .164.$$

The odds of this person donating are 0.16 to 1 or 16 out of 100—not very high. Remember these are odds, not probabilities. Compare these odds with a 45-year-old person one income level up:

$$\hat{O} = e^{-2.89}e^{.02(45)}e^{.18(2)} = e^{-1.63} = .196.$$

The odds have increased slightly, a result consistent with the positive sign on the coefficient for income. Finally, look at a 45-year-old in the highest income level (income = 11):

$$\hat{O} = e^{-2.89}e^{.02(45)}e^{.18(11)} = e^{.01} = .990.$$

We can proceed in this manner to find all the predicted odds for income = 1, . . . , 11 (see table 14-24).

Note two things. First, as income increases, so do the odds of making a donation. At all levels, they are less than 1, which means that no matter what their income people are more likely not to give than to give. Second, the differences in the odds are not constant. Moving from one category to the next sometimes produces a miniscule change in the odds of giving; sometimes, the change is larger. That's why we say the effects of income (and age) are multiplicative, not additive. (If they were additive, the difference in odds would be constant as we moved from one level to the next.) For instance, the coefficient for income is 0.18; if you obtain its exponent

TABLE 14-24 Predicted Odds of Political Donation by Income Level

Income level	Probability of Donating	Odds of Donating	Change in Odds
1	.126	.164	–
2	.147	.196	.032
3	.172	.235	.039
4	.200	.281	.046
5	.231	.336	.055
6	.265	.403	.067
7	.303	.482	.079
8	.343	.577	.095
9	.386	.691	.114
10	.430	.827	.136
11	.476	.990	.163

Change = difference between current odds and odds one level above (e.g., .173 – .144 = .029).

Note: Results subject to rounding errors.

($exp(.18) \approx 1.20$) and multiply it by the odds at a given income level, you obtain the odds for the next level.

We should stress that these remarks are simply an alternative but equivalent way of interpreting logistic regression coefficients. Moreover, we can move from one view to the other by simply manipulating the results with a pocket calculator. Most computer programs and articles report the coefficients, along with other statistical information. To make sense of them often requires substituting actual data values into the equations and seeing what the probabilities or odds turn out to be.

HELPFUL HINTS

Probability versus Odds

Keep the terms straight. A probability is not the same as odds, at least in statistical analysis. A probability refers to the chances of something happening, such as a person donating money to a cause. Odds compare two probabilities, such as the probability of contributing to the probability of not contributing. If N_Y is the number of people out of a sample of N who give, for example, the estimated probability of giving is

$$\hat{P}=\frac{N_Y}{N}.$$

The estimated probability of a "no" is

$$\hat{Q}=1-\hat{P}=\frac{N-N_Y}{N}.$$

The estimated odds of observing a "yes" as opposed to a "no," however, are

$$\hat{Q}=\frac{\hat{P}}{\hat{Q}}=\frac{\frac{N_Y}{N}}{\frac{N-N_Y}{N}}=\frac{N_Y}{N-N_Y}.$$

If the probability of donation is .6, then the probability of not is 1 − .6 = .4, and the corresponding odds are .6/.4 = 1.5 or 1.5 to 1.

Check out more Helpful Hints at **edge.sagepub.com/johnson8e**

Logits

Finally, for the sake of completeness, we note that a logistic regression model can also be expressed as a linear function of the variables, but the dependent variable is then the natural logarithm of the odds, a quantity that many practitioners find difficult to grasp in a substantive context. Still, the *log* odds (logit) can be expressed as a linear model in which the coefficients have their usual meaning: holding other independent variables constant, a one-unit increase in *X* leads to a $\hat{\beta}_x$ beta change in the logit. Suppose we fix age at 45 and compare logits of people at income levels 5 and 6 (1 unit apart):

$$\text{Logit}_{\text{Income=5}} = -2.89 + .02(45) + .18(5) = -1.09;$$
$$\text{Logit}_{\text{Income=6}} = -2.89 + .02(45) + .18(6) = -.91;$$
$$\text{Difference} = (-1.09) - (-.91) = -.18 = \hat{\beta}_{\text{Income}}.$$

Clearly, changing income one place increases or decreases the log odds of donating by .18. But how does one make theoretical or practical sense of a logit? Its lack of a clear meaning leads many analysts to convert the logit to odds by exponentiating.

Conclusion

As we have seen, multivariate data analysis helps researchers provide more complete explanations of political phenomena. Observing the relationship between an independent and a dependent variable while controlling for one or more control variables allows researchers to assess more precisely the effect attributable to each independent variable and to accumulate evidence in support of a causal claim. Being able to observe simultaneously the relationship between many independent variables and a dependent variable also helps researchers construct more parsimonious and complete explanations for political phenomena.

Multivariate data analysis techniques control for variables in different ways. Multivariate cross-tabulations control by grouping similar observations; partial correlation and multiple and logistic regression control by adjustment. Both types of procedures have their advantages and limitations. Control by grouping can result in the proliferation of analysis tables, the reduction of the number of cases within categories to a problematic level, and the elimination of some of the variance in the control variables. Control by adjustment, in contrast, can disguise important aspects of relationships—that is, relationships that are not identical across the range of values observed in the control variables.

TERMS INTRODUCED

Control by grouping. A form of statistical control in which observations identical or similar to the control variable are grouped together.

Dummy variable. A hypothetical index that has just two values: 0 for the presence (or absence) of a factor and 1 for its absence (or presence).

Interaction. The strength and direction of a relationship depend on an additional variable or variables.

Logistic regression. A nonlinear regression model that relates a set of explanatory variables to a dichotomous dependent variable.

Logistic regression coefficient. A multiple regression coefficient based on the logistic model.

Multiple regression analysis. A technique for measuring the mathematical relationships between more than one independent variable and a dependent variable while controlling for all other independent variables in the equation.

Multiple regression coefficient. A number that tells how much Y will change for a one-unit change in a particular independent variable, if all the other variables in the model have been held constant.

Multivariate cross-tabulation. A procedure by which cross-tabulation is used to control for a third variable.

Partial regression coefficient. A number that indicates how much a dependent variable would change if an independent variable changed one unit and all other variables in the equation or model were held constant.

Regression constant. Value of the dependent variable when all the values of the independent variables in the equation equal zero.

SUGGESTED READINGS

Agresti, Alan, and Barbara Finlay. *Statistical Methods for the Social Sciences.* 3rd ed. Saddle River, N.J.: Prentice Hall, 1997.

Anderson, T. W. *Introduction to Multivariate Statistical Analysis.* 3rd ed. New York: Wiley, 2003.

Berk, Richard A. *Regression Analysis: A Constructive Critique.* Thousand Oaks, Calif.: Sage, 2004.

Blalock, Hubert M., Jr. *Causal Inference in Non-Experimental Research.* Chapel Hill: University of North Carolina Press, 1964.

Draper, Norman R., and Harry Smith. *Applied Regression Analysis.* 3rd ed. New York: Wiley, 1998.

Fox, John. *Applied Regression Analysis and Generalized Linear Models.* 2nd ed. Los Angeles: Sage, 2008.

Long, J. Scott. *Regression Models for Categorical and Limited Dependent Variables.* Thousand Oaks, Calif.: Sage, 1997.

Overall, John. E., and C. James Klett. *Applied Multivariate Analysis.* New York: McGraw-Hill, 1973.

Pampel, Fred C. *Logistic Regression: A Primer.* Series: Quantitative Applications in the Social Sciences, 132. Thousand Oaks, Calif.: Sage, 2000.

CHAPTER 15

The Research Report: An Annotated Example

Annotated Research Report Example

Predicting Presence at the Intersections: Assessing the Variation in Women's Office Holding across the States

Becki Scola[*]

Abstract

Over the past several decades, women's office holding at the state level has grown substantially, but there is still a large range of electoral service across the 50 states. In this article, I revisit the most common explanations provided by the literature in helping us understand this variation and assess whether these explanations can be effectively applied to different racial/ethnic groups of female legislators. Using data from a 20-year time span, I find that there are differences between the factors that predict white women and women of color's state legislative presence.

Keywords

Identity/group politics, gender politics, political culture, representation, electoral systems

Introduction

In 1990, two years before the "Year of the Woman" was heralded as a new age for women in politics, women represented 17.1% of all state legislators across the nation (Center for American Women and Politics [CAWP] 2005). By 2010, 24.5% of all state legislators were women (CAWP 2010a). While the descriptive representation of women did indeed increase over this 20-year time period, the aggregate percentages veil a more complex picture of women's legislative service. One such complexity is the variation we see in the percentage of female office holding across states. For example, in 2010, 38% of Colorado's legislature comprised women, while South Carolina's consisted of only 10%. This is not a new trend—we have consistently witnessed a range in female state legislative office holding across the states for the past four decades (Norrander and Wilcox 1998).[1]

Nor is it new to investigate the factors that might explain women's legislative office holding. Previous studies suggest that women are more likely to serve in states that have larger pools of potential candidates (Arceneaux 2001; Camobreco and Barnello 2003; Hogan 2001; Nechemias 1987; Norrander and Wilcox 1998; 2005; Rule 1990; 1999), liberal political ideologies (Arceneaux 2001; Camobreco and Barnello 2003; Norrander and Wilcox 1998; 2005), and Moralistic political cultures (Arceneaux 2001; Camobreco and Barnello 2003; Hill 1981; Hogan 2001; Nechemias 1987; Norrander and Wilcox 1998). There are mixed results for states with more professional legislatures, with studies noting that the predictive capacity of these variables depends on the time frame and context.

Over this same time period, women of color's presence in state legislatures increased from 1.8% in 1990 to 4.8% in 2010, with American Indian women serving in as few as seven legislatures to black women serving in as many as 38 state legislatures (CAWP 2010b). We also saw a variation in their office holding across the states in 2010, from a high of 22.4% in Hawaii to a low of 0% in six states. Table 1 not only illustrates the cross-sectional variation we see for female office holding in general for 2010 but also highlights the differences between white female and women of color legislators in terms of this variation. Does the conventional wisdom apply to the variation we see in women of color's legislative presence? This is still an open question.

Question Set 1: Here is where the author begins to identify her research question: Are the factors that explain variation in female office holding among the states the same for women of color?

[1] In 1990, the percentages of women in state legislative office ranged from a low of 2.1% in Louisiana to a high of 33.3% in Vermont. In 1995, the range was between 3.6% (Alabama) and 39.5% (Washington). In the year 2000, Alabama and Washington again ranked as the lowest and highest states, with 7.9% and 40.8%, respectively. In 2005, South Carolina held the bottom spot with 8.2%, while Maryland rang in at Number 1 with 35.6%. By 2010, Colorado took the Number 1 spot with 38%, and South Carolina remained at the bottom with 10%.

TABLE 1 **Percentage of Women, White Women, and Women of Color in State Legislature, 2010.**

State	Total in Legislature	Women (%)	White Women (%)	Women of Color (%)
Alabama	140	12.9	5.7	7.1
Alaska	60	21.7	20.0	1.7
Arizona	90	32.2	18.9	13.3
Arkansas	135	23.0	20.7	2.2
California	120	26.7	16.7	10.0
Colorado	100	38.0	30.0	8.0
Connecticut	187	32.1	28.9	3.2
Delaware	62	25.8	22.6	3.2
Florida	160	23.8	15.0	8.8
Georgia	236	19.5	8.5	11.0
Hawaii	76	32.9	3.9	28.9
Idaho	126	25.7	19.0	2.4
Illinois	177	28.2	16.9	11.3
Indiana	150	21.3	18.0	3.3
Iowa	150	23.3	20.7	2.7
Kansas	165	30.3	26.7	3.6
Kentucky	138	15.9	15.9	0.0
Louisiana	144	16.0	10.4	5.6
Maine	186	29.0	29.0	0.0
Maryland	188	31.4	18.1	13.3
Massachusetts	200	25.5	23.0	2.5
Michigan	148	25.0	22.3	2.7
Minnesota	201	34.8	33.3	1.5
Mississippi	174	14.4	6.9	7.5
Missouri	197	22.3	16.8	5.6

State	Total in Legislature	Women (%)	White Women (%)	Women of Color (%)
Montana	150	26.0	22.7	3.3
Nebraska	49	20.4	18.4	2.0
Nevada	63	31.7	22.2	9.5
New Hampshire	424	36.8	36.3	0.5
New Jersey	120	28.3	15.8	12.5
New Mexico	112	30.4	15.2	15.2
New York	211	24.1	17.5	6.6
North Carolina	170	25.9	18.2	7.6
North Dakota	159	16.3	14.5	0.0
Ohio	132	22.0	15.9	6.1
Oklahoma	149	11.4	9.4	2.0
Oregon	90	28.9	26.7	2.2
Pennsylvania	253	15.4	11.9	3.6
Rhode Island	150	22.1	15.3	1.3
South Carolina	170	10.0	7.1	2.9
South Dakota	105	20.0	20.0	0.0
Tennessee	132	18.9	12.9	6.1
Texas	181	23.8	12.7	11.0
Utah	104	22.1	19.2	2.9
Vermont	180	37.2	36.7	0.6
Virginia	140	19.3	12.9	6.4
Washington	147	32.7	30.6	2.0
West Virginia	134	16.4	14.9	1.5
Wisconsin	132	22.0	18.9	3.0
Wyoming	90	16.7	15.6	1.1

in state legislatures (Arceneaux 2001; Hill 1981; Hogan 2001; Nechemias 1987; Norrander and Wilcox 1998; 2005).

Highly professionalized legislatures (those with high salaries, large staffs, and longer sessions) might make a legislative career more attractive, which in turn may generate an increase in male candidates, putting potential female hopefuls at a disadvantage (Arceneaux 2001; Hill 1981; Hogan 2001). Then again, men, on average, have higher incomes than women in the private sector, which suggests that men encounter a higher opportunity cost when they run for office. Nevertheless, the literature supports the idea that we will see fewer women in more professional legislatures (Arceneaux 2001; Hill 1981; Hogan 2001; Nechemias 1987), although this finding seems to depend on the context and time frame under investigation.

Question Set 3: In these paragraphs, the author succinctly identifies previous studies and summarizes their findings. Notice that her discussion of these studies is organized around the factors found to explain variation in the presence of women in state legislatures.

Predicting Presence at the Intersections?

The model derived from earlier work on the cross-sectional variation of women in state legislatures provides us with five consistently used, comparable factors that best predict the presence of female office holders: pool of candidates, political culture, political ideology, minority population, and legislative professionalization. Do these predictors differ if we take the race/ethnicity of the female legislator into account? In other words, does the model explain the variation we see in both white women and women of color's office holding? The theory of intersecting identities provides us with a good reason to suspect that the model may not be as useful in helping us understand the range in the sex composition of state legislatures if we disaggregate by race/ethnicity.

Intersectionality theorists contend that we have multiple identities that inform political activity. We are not "women" or "minorities"—we are both simultaneously (Crenshaw 1991; 1989). Therefore, scholars cannot fully describe, study, and understand political phenomenon without addressing the multiple identities that give rise to these experiences. To be sure, several women and politics scholars have employed an intersectional framework to study how race and gender intersect at the elite level in terms of background characteristics (Darling 1998; Fraga et al. 2003; Hardy-Fanta et al. 2007; Moncrief, Thompson, and Schuhmann 1991; Prestage 1991; Williams 2001) and political ambition (Darcy and Hadley 1988; Hardy-Fanta et al. 2007), as well as at the mass-public level in terms of public opinion (Gay and Tate 1998) and candidate support (Philpot and Walton 2007). Other studies have documented how race intersects with gender at the judicial level (Collins and Moyer 2008) and in public policy making (Fraga et al.

2005; Prindeville and Gomez 1999; Smooth 2001). Smooth (2006) applies the theory to the context of the Voting Rights Act and highlights how our traditional understanding of electoral politics changes when viewed through an intersectional lens. Hawkesworth (2003, 530) contends that Congress itself is a "racedgendered institution" that produces "raced and gendered hierarchies that structure interactions among member as well as institutional practices." She suggests further exploring the within-group differences among women so as to more fully explain the processes that create and maintain both gender and racial hierarchies.

Hence, the theory of intersectionality might lead us to expect that the factors cited as most important in predicting female presence within state legislatures will vary based on the within-group difference of race/ethnicity. From this theoretical perspective, I hypothesize that the indicators most commonly used in explaining the variation in women's office holding across the states will not perform equally in predicting both white women and women of color's range of legislative service. Disaggregating female legislators by race/ethnicity highlights how race intersects with gender as a politically relevant characteristic. Indeed, there is evidence that these intersecting identities matter at the congressional level (Palmer and Simon 2008). Palmer and Simon (2008) document that the congressional districts that elect white women and black women are distinct with regard to socio-demographic and institutional characteristics.

Question Set 3: This section reviews the literature related to the theory of intersectionality. It is not clear, however, why intersectionality is referred to as a "theory," but the central idea is that racial and gender identity need to be considered together for a better understanding of political phenomena.

An excellent example of how the intersectional approach adds nuance to our analyses is the case of term limits. Theoretically, terms limits should positively influence the election of women and minorities since the incumbency advantage significantly contributes to the underrepresentation of both women and minorities. Incumbents are primarily white males, so the removal of this barrier should support the election of underrepresented groups. Research on whether term limits assist or hinder women is at odds, though, with most studies concluding that term limits may help women only after they are first implemented—the effect seems to diminish the longer they have been in place (Caress 1999; Carroll and Jenkins 2001a; 2001c).

Of interest here is the intersectional effect of term limits. Carroll and Jenkins (2001b, 8) note that in 1998, women lost, minorities gained, and women of color "more closely resemble the patterns for women than... minorities." Furthermore, they carefully note that this pattern varies by race/ethnicity and by year: black women lost seats and Latina gained seats. In contrast, for the 2000 election, "the pattern for minority women parallels the pattern for minorities more generally" (Carroll and Jenkins 2001b, 8) for both black women and Latinas. The study of how term limits impact female office holding are frequently included in studies that look at the increase in female representation over time and not typically present in projects that look directly at the variation. In other words, since the current investigation is grounded in the variation literature and does not attempt to explain change over time in women's office holding, term limits are not included in the analysis. Nevertheless, this intersecting illustration of term limits provides leverage to the idea that the five variables under investigation in this study might vary as well. It is to these factors that I now turn and offer some expectations from an intersectional framework.

Question Sets 4 and 6: Here the author presents her first hypothesis. It is not clear at this point how she intends to measure the independent variable, the size of the potential pool of candidates, and how this concept will be distinct from the educational or income levels of women. Nevertheless, this hypothesis and subsequent hypotheses are empirical, general, and plausible.

The potential pool of candidates consistently predicts female office holding across the states. I hypothesize that the current study will confirm the positive and significant effect of this variable for predicting female legislative service, but that the results will be greater for women of color legislators than white women. The primary reason to anticipate a stronger influence for women of color is based on findings from previous research, which notes that women of color legislators have higher levels of education than do their white female counterparts (Moncrief, Thompson, and Schuhmann 1991). This suggests that the pool of candidates from which women of color emerge is perhaps more affluent when compared directly with the pools of white women at the state level.

Question Set 4: This is the second hypothesis in which the political culture of a state is the independent variable. The author hypothesizes that the Moralistic category of political culture is associated with only white women's legislative service. The author notes that political culture also varies with region.

Question Set 6: The author does not explain the connection between political culture and women's legislative service. Her hypothesis is justified on the basis that previous research has found such a connection.

Question Set 4: Here the author hypothesizes that states with liberal political ideologies will have higher levels of white women legislators.

Women of color legislators are concentrated mainly in the southern and western regions of the United States—states that are generally designated as Traditionalistic and Individualistic. White women's office holding is more prominent in the northeastern and midwestern United States—states that are typically designated as Moralistic. Hence, my expectation is that states that are identified as having a Moralistic political culture will positively and significantly relate to white women's legislative service, but will negatively predict women of color's service in state legislatures. Furthermore, states that are characterized as Traditionalistic are also states where we typically see higher levels of conservative political ideology. Following the same reasoning from above in terms of the concentration of white women and women of color legislators, I believe my test will confirm the value of higher levels of liberal political ideologies in predicting where white women will serve, but that we will not see a similar relationship for women of color legislators.

Relatedly, higher percentages of minority populations are concentrated in certain regions of the nation and map onto the pattern of minority office holding fairly well—where we see higher percentages of minority populations, we also see higher percentages of minority office holders (Hardy-Fanta et al. 2005). My expectation is that higher percentages of minority populations will be positively and significantly related to women of color's office holding. Clearly, a substantial minority population is a necessary condition for electoral service, both in terms of how it affects the potential pool of candidates and with regard to the minority electorate's desire to elect "one of their own" (Philpot and Walton 2007). For white women legislators, I expect the relationship to be positive but not significant. I am not suggesting that higher percentages of minority populations will negatively impact white women's legislative service. What I suspect is that previous tests of the model may be capturing women of color's legislative presence, and this particular variable may not have the same explanatory power within an intersectional framework.

Question Set 4: The fifth hypothesized relationship predicts that higher percentages of minority populations will lead to higher levels of women legislators, but the author predicts that the relationship will be stronger for women of color than for white women.

The author notes that political culture, political ideology, and percentage of minority populations are all connected to region. Rather than hypothesize that region predicts the presence of women in state legislatures, the author identifies the factors that account for the connection between region and women legislators.

The level of legislative professionalization presents a unique opportunity for applying an intersectional framework. The research that looks at how professionalization affects women and minorities tends to find a negative relationship for women and positive relationship for minority office holding (Hero 1998; Squire 1992). How professionalization relates to women of color legislators is not addressed. Thus, I will rely on the logic presented in the variation literature, which assumes that more professionalized legislatures make office holding more appealing and, therefore, more competitive. Since female candidates may be less likely to run in these competitive environments (Lawless and Fox 2005), I speculate that professionalization will negatively affect the legislative service of both white women and women of color.

Question Set 4: The author hypothesizes that women's legislative service depends on the level of professionalization of a state's legislature.

[3] Data collected for this project was made possible in part by Summer Research Grants from the School of Social Sciences and the Department of Political Science at University of California, Irvine.

[4] The selection of the data points represents five-year intervals. The year 1990 marks a point in time in which we begin to see structural changes in state legislatures, for instance, increasing professionalization. Other years offer differing electoral contexts (i.e. 1995 represents a point in time after the 1992 "Year of the Woman" elections; 2000 was presidential election year; 2005 represents office holding after the 2000 census and, thus, redistricting changes; and the year 2010 is the most recent year for which I could collect comparable data).

Question Set 7: In this section the author describes the operational definitions of her variables. Each of the operational definitions appears valid and reasonable. She does not explicitly mention the level of measurement of each variable. Most of the variables are percentages, which would be ratio-level measures. She uses several scale measures (to measure political culture, professionalization of the state legislature, and political ideology), which she acceptably treats as interval or ratio -level measures.

Question Set 8: The author obtains her data from documents from a variety of sources, including the Census Bureau, the Center for American Women and Politics, and the work of other political scientists who classified states with respect to their political culture, political ideology, professionalization of the state legislature, and party control.

Question Set 9: The unit of analysis is the state. Each of the variables measures an attribute of states.

Question Set 10: The author does not use a sample. She includes the entire population of states in her analysis.

Question Sets 5 and 7: The author clearly designates three dependent variables: the percentage of women in a state's legislature, the percent of white women, and the percentage of women of color, and explains how they will be measured.

The author does not explain why she calculated these percentages for each of the states for the five time periods, which she then averaged to obtain a single measure for each of the three dependent variables. One reason for doing this is to improve the validity of these measures. If she had used only one year for her analysis, it is possible that particular year would not have been the most typical year for one or more states.

In addition, several of the measurements of independent variables cover a longer period of time than a single year.

Testing the Model at the Intersections: Data and Method

The question guiding this study pertains to the cross-sectional variation in women legislators across the states. Previous work on the topic suggests that there is a set of variables that perform fairly well in explaining this variation. My goal is to test the effectiveness of these factors from an intersectional perspective. Accordingly, I collected data that included the percentage of women serving in all 50 state legislatures, disaggregated by race/ethnicity, as well as several state-level demographic and legislative indicators for a 20-year time span (1990–2010).[3] Three dependent variables were constructed: the percentage of women in the state legislature, the percentage of white women, and the percentage of women of color (African American, Latina, Asian and Pacific Islander, and American Indian). All three dependent variables were calculated as a pooled percentage of the total legislature for each of the 50 states for five points in time (1990, 1995, 2000, 2005, and 2010),[4] where the percentages for each of the five time periods were averaged, and the mean from this calculation became the three dependent variables. Information on the race/ethnicity of the female state legislators was gathered from the CAWP.

Closely following the previous literature, I identified five independent variables that were the most common across studies that investigated the variation in female office holding and that had also significantly predicted the presence of female legislators: pool of potential candidates, political culture, percentage of minority population, political ideology, and professionalization of the state legislature. For purposes of comparability, the independent variables were measured in much the same way that other variation studies operationalize these concepts with two exceptions. First, the “pool of potential candidates’” variable typically consists of two measures: the percentage of women with higher education and the percentage of women in the workforce. To overcome any potential correlational problems associated with these two indicators, my measure for this variable was the percentage of professional women in the state as identified by the U.S. Census Bureau (1990; 2000) Summary Files. Arguably, a woman who has a professional career has a higher education *and* is in the workforce, thereby capturing the effect of both of these indicators in one variable.

Question Sets 5 and 7: Here the author explains how each of the independent variables was measured.

Second, political culture is cited as one of the most significant indicators for explaining the cross-sectional variation in female office holding. While all the comparable studies attest to its predictive capacity, there is no consistent measurement of the variable. Some studies use Elazar's ([1984] 1966) categories of Moralistic and/or Traditionalistic cultures as dummy variables (Arceneaux 2001; Hogan 2001), others use Sharkansky's (1969) scale (Nechemias 1987; Norrander and Wilcox 1998; 2005), and some use Johnson's (1976) scale (Hill 1981; Nechemias 1987). Elazar's ([1984] 1966) original typology categorized states as Moralistic, Traditionalistic, Individualistic, or a combination of two of these. To capture the mixture designations, Sharkansky developed a scale ranging from 1 (*purely Moralistic*) to 9 (*purely Traditionalistic*). Johnson developed a subculture index based on Elazar's designations through a measure of religiosity. I selected Sharkansky's scale to measure political culture for this project since his index includes "mixed" cultures, which allow for a more nuanced analysis of Elazar's categories and the variations of political culture designations. If previous studies are correct, we should expect to see fewer female state legislators the closer a state gets to 9 on the scale.

The measurement of the remaining independent variables runs parallel to the other variation studies. The minority population was calculated as a percentage of the nonwhite state population (black, Latino, Asian American, Pacific Islanders, and Native Americans) as indicated by U.S. Census Bureau data. Political ideology was measured using Erikson, Wright, and McIver's (2007) files, which are now appropriately updated for state-level and multiyear application. They used a 0-to-1 scale, with "0" indicating *more conservative* and "1" indicating *more liberal*. I applied Squire's (2007) scale for measuring the professionalization of the state legislature, which accounts for the time involved, resources available, and salary of state legislators. The 0 (*least professional*) to 1 (*most professional*) scale is included in this analysis. Finally, since a majority of female legislators (and more than 85% of women of color legislators) are Democrats, I included a control variable for party control based on Ceaser and Saldin's (2005, 247) "major party index," which "is intended to measure the level and extent of interparty competition in and between the states" and "is comprised of six weighted components calculated on even numbered years for each state from 1990 to 2002."

Question Set 5: The author includes state party competition or party control as a control variable. Because most female legislators are Democrats, she needs to hold party control constant to see if the other variables influence the election of women to the state legislature independently of whether the Democratic Party dominates a state's politics.

There are a few limitations to this study that should be noted. First, the data for this project consist of aggregate-level data. This has three implications. One, it does not differentiate among the 50 states. Since the same model may not be appropriate for all states, a state-by-state analysis may uncover substantial variations in the patterns we see here. Two, it comprises state-level data that do not account for district-specific characteristics. District level data might provide a more nuanced analysis, especially for variables such as the percentage of the minority population. Three, I collapse both chambers of the legislature. This may mask differences between the senate and house. Nonetheless, the intention here is to ascertain a general pattern, and the method and data used are appropriate to the task and not unlike other studies of its kind.

Question Set 11: The author draws our attention to limitations in the chosen research design, which is a cross-sectional design. It is misleading, however, to say that her model does not differentiate among the fifty states. It should be relatively easy to determine for which states the model fits best, and for which states the predictions are not a good fit. A greater concern involves the fact that the measures for the independent variables are collected at the state level, although legislators are elected by district and districts may vary considerably with respect to values of the independent variables. A third limitation is that the analysis is for state legislatures as a whole, not disaggregated by legislative chamber, and the factors that influence the election of women to the lower chamber may be different from the factors that elect women to the upper chamber.

Another important limitation is that the dependent variable for women of color legislators combines all racial/ethnic groups. Undoubtedly, combining all minority women might obscure what is truly going on. Separate models for black women, Latinas, Asian American women, and American Indian women would provide a richer story of how race intersects with office holding, and better inform us about the nuances and differences that exist among racial and ethnic groups of women. That being said, there is some precedent for using the term *women of color* and for collapsing data. Hawkesworth (2003) invokes this term in her study of congressional women of color and combines the data she gathered for African American, Latina, and Asian American women. Hawkesworth's (2003, 532) reason for doing so is the same as mine: the "small n." While this justification may not be wholly satisfying (for the reader or the author), collapsing the data was necessary to have a sufficient number of cases to test the model.

Question Set 2: Earlier, the author discussed how her research would contribute to our understanding of the election of women and women of color to state legislatures, but her research doesn't address all possible questions about the interaction between race and gender. One limitation of the study is that all women legislators of color are combined. Thus, the comparisons are between white female legislators and those of color. While it might be interesting to compare women of several different races or ethnicities, there are simply not enough women legislators of these different groups to analyze statistically.

Results and Discussion: Women (of Color) and Legislative Presence

Table 2 presents the ordinary least squares (OLS) regression coefficients (and standard errors) for the percentage of women, white women, and women of color in state legislatures. The first thing to note is that the model is statistically significant and performs fairly well (0.491 adjusted R^2) in explaining the variation we

TABLE 2 **Unstandardized Regression Coefficients (and Standard Errors) for the Percentage of Women, White Women, and Women of Color in State Legislatures.**

	Women	White Women	Women of Color
Percent professional women	2.413* (0.064)	2.163* (1.267)	0.259 (0.708)
Political culture	−0.010** (0.004)	−0.013*** (0.004)	0.004* (0.002)
Liberal ideology	0.828*** (0.307)	0.672** (0.303)	0.176 (0.169)
Percent minority population	−0.038 (0.089)	−0.203** (0.087)	0.161*** (0.049)
Legislative professionalization	−0.099 (0.064)	−0.152** (0.063)	0.055 (0.035)
Control: rep party control	0.135 (0.111)	0.133 (0.109)	0.008 (0.061)
Constant	−0.018 (0.137)	0.044 (0.105)	−0.072 (0.059)
R^2	0.491	0.582	0.431
Significance of model (*F*-statistic)	8.882***	12.370***	7.177***
N	50	50	50

*p <.10. **p <.05. ***p <.01

see in women's legislative service. In terms of our variables of interest, three expectations are confirmed: states with higher percentages of professional women, higher levels of liberal political ideologies, and those that come closer to approximating a Moralistic political culture are all significant for positively predicting higher percentages of female state legislators.

Against my expectation and the previous literature, a higher percentage of minorities within the population does not significantly predict female service. One possible reason for this could be that this variable is operationalized using aggregate state-level data. As stated earlier, measuring the minority population in this fashion may be hiding district-level characteristics. Legislative professionalization was a negative indicator of women's service but not a significant indicator for explaining the variation in female office holding.

Question Set 12: The statistics used to analyze the data are clearly stated and are appropriate for the level of measurement of the variables.

Question Set 13: The results of the study are clearly presented in table 2. The author uses R^2 to measure the overall fit of the model and the *F*-statistic to indicate the statistical significance of the model. She reports the regression coefficients and the standard errors for each of the independent variables. She reports their statistical significance using the asterisk (*) method.

Question Set 13: The author predicted that the same set of independent variables would not be equally useful in predicting the percentage of white female legislators and the percentage of female legislators of color, and her results support this prediction. The model predicts the percentage of white women legislators better as indicated by the R^2 values.

How does the model perform when we apply an intersectional framework? I argued earlier that if we take the race/ethnicity of the legislator into account, the factors in the model would not offer the same predictive capacity as cited in previous studies. Table 2 offers some initial support for this hypothesis if we consider the percentage of white women legislators and the percentage of women of color legislators for the same set of independent variables. While both models reach statistical significance, the model better explains the variance for white women (0.582 adjusted R^2) than that for women of color (0.431 adjusted R^2).

Question Set 13: Here the author focuses on her predictions regarding the election of white women legislators. The variables that explained the variation in the percentage of all female legislators are also the ones that are statistically significant in explaining variation in the percentage of white female legislators. In addition to these variables, another independent variable, the percentage of minority population in a state, is statistically significant in predicting the presence of white women legislators. The sign of the coefficient, however, is in the opposite direction of what the author had predicted.

Assessing the results for white female legislators, most of the expectations are confirmed: states with higher levels of professional women, liberal ideologies, and Moralistic political cultures have higher percentages of white women legislators, while states with higher levels of professionalization have lower percentages. The result for the percentage of the minority population not only goes against what I expected but also does not comport with the previous literature—it is negatively related to white female service. I offer a possible explanation for this result below when I discuss women of color's service. Overall, three of the variables that best explain white female's presence in legislatures are similar in direction and significance to the variables that help us understand the cross-sectional variation in women's legislative service.

Applying the same model to women of color's office holding, we see that the predictive capacity and direction of the variables in the model changes substantially. Two expectations are confirmed: states with higher percentages of minority populations and states that fall closer to *Traditionalistic* political cultures tend to have higher percentages of women of color legislators. I originally hypothesized that the percentage of professional women in the state would matter more for women of color than for white women. This is not the case. Even though the percentage of professional women in the state has a positive effect on women of color's legislative service, it is not a significant relationship. This might be related to how women of color enter the political arena, with some studies suggesting that their path differs from that of white women (Lawless and Fox 2005; Moncrief, Thompson, and Schuhmann 1991; Williams 2001). Liberal political ideologies, as expected, do not have an independent effect, nor does the level of legislative professionalization in explaining the cross-sectional variation in women of color's presence in state legislatures.

Recall that when we do not take the race/ethnicity into account, the percentage of the minority population does not perform as expected. However, when we disaggregate female legislators by race/ethnicity, this factor is negatively associated with white female legislators and positively associated with women of color legislators. This makes intuitive sense: women of color are more likely to emerge in states with higher percentages of people of color, while this does not have an independent effect on where we see white women emerge. Two things are particularly interesting about these findings. First, the extant literature consistently confirms that higher percentages of minority populations predict higher percentages of female legislators. Perhaps this is a function of the time frames under consideration in the other studies—most were conducted at a time when we see women of color's legislative presence increasing, and they could be capturing this phenomenon. Since the data for this project moves us into the 2000s, when all of women's legislative service leveled off, it is possible that my assessment notes a declining importance of this factor. Theoretically, though, the findings discussed here may confirm not the declining importance per se, but the *relative* importance of this particular variable. Indeed, higher percentages of nonwhite populations positively and significantly predict the presence of women of color legislators, but negatively and significantly predict white females.

Question Set 13: In these paragraphs the author focuses on whether the results support her hypotheses regarding the factors associated with the election of women of color to state legislatures. For women of color, only two independent variables are statistically significant—political culture and percent minority population. But for women of color, the predicted association between Moralistic political culture and the presence of women legislators is not supported by the data. Instead, the presence of women of color in state legislatures is associated with the Traditionalistic political culture. This actually makes sense because southern states are predominantly Traditionalistic and also have higher percentages of minority populations.

The other item that goes against the previous literature but confirms my expectation is the political culture variable. Again, Moralistic political culture is one of the most reliable factors for explaining cross-sectional variation, and it does perform as expected for female legislators overall as well as for white women in particular. This pattern does not hold for women of color legislators. In fact, we are more likely to see their presence as states move closer to the Traditionalistic end of the culture scale.

On one hand, a possible explanation for this finding might be related to region. Almost all southern states have Traditionalistic political cultures, and these are also the states where we see higher percentages of women of color in the legislature. Higher levels of white women, conversely, serve in northeastern and midwestern states, which are more likely to have Moralistic political cultures.

On the other hand, both white women and women of color have higher levels of legislative service in western states, which are typically categorized as a mix of Individualistic/Moralistic or Individualistic/Traditionalistic. Furthermore, Elazar's typology is in part based on region (namely by attempting to explain regional differences), and Sharkansky partially controls for the inclusion of Individualistic and "mixed" states with his scaled estimations. Region, then, may not be the only mitigating factor when interpreting these results.

Conclusion: Predicting Presence at the Intersection

The purpose of this article was to revisit the explanation for the variation we see in women's legislative service across the state from an intersectional framework. Applying this theoretical concept to the set of variables that have been consistently used to explain female legislator's cross-sectional variation, I tested the model's predicting capacity to determine whether these variables were the same or different when comparing white women and women of color. My analysis offers some initial support for the theory that intersection matters with regard to female office holding.

The explanatory power of five independent factors under consideration here varies depending on whether female legislators are white or of color. All in all, white women and women of color legislators had very little in common in terms of predictors. We are more likely to see higher percentages of white female legislators in states with higher percentages of professional women, higher levels of liberal political ideologies among the electorate, and those classified as having a Moralistic culture. We are less likely to see white female legislators in states with higher percentages of minority populations and professionalized legislatures. Women of color legislators are more likely to serve in states that fall closer to having a Traditionalistic political culture and higher percentages of minority population.

Question Set 13: This is a clear summary of the research findings.

The key idea that I would like to emphasize is that the model for explaining female legislative service more closely matches the white women legislator's level of service than women of color's office holding. This indicates that previous findings more accurately describe white women legislator's presence and less accurately predict the proportion of women of color state legislators. Instead, the research presented here reveals that the indicators most useful in predicting women's legislative service are noticeably raced, as they are less helpful in explaining where women of color serve. Clearly, race and gender are intersecting when it comes to legislative office holding. In short, race/ ethnicity did make a difference when we applied the model to different groups of female legislators.

In conclusion, my findings are merely a stepping stone for future research. Exactly *how* and *why* gender and race/ethnicity are intersecting is underexplored in terms of legislative office holding. To be sure, the process seems to be more complex than what is captured by race/ethnicity *or* gender separately. While I have imposed the same (perhaps rigid) model on all women of color for reasons of parsimony, the evidence suggests that this singular model does not help us explain all of the variance we see in state legislative office holding. Of course, collapsing all minority women into one dependent variable may obscure what is truly going on. Indeed, running separate models for African American women, Latinas, Asian American, and Native American women, as data permit, would provide a richer story of how gender, race, and ethnicity influence legislative service.

If we are to better understand how this intersection functions in terms of legislative representation, we need to begin pulling apart the traditional models and replacing them with intersecting models of legislative office holding. As we move forward, it is in our best interest to ask, "Why, where, when, and how do each of the variables matter, and for whom?"

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Note: It is now common practice for authors to reveal whether they might benefit personally by their research findings and also to reveal any outside sources of funding or sponsors. In some situations, funding sources may have a preference or particular interest in research findings. These two statements allow consumers of the research to assess better its objectivity.

References

Arceneaux, Kevin. 2001. "The 'Gender Gap' in State Legislative Representation: New Data to Tackle and Old Question." *Political Research Quarterly* 54: 143–60.

Camobreco, John F., and Michelle A. Barnello. 2003. "Postmaterialism and Post-Industrialism: Cultural Influences on Female Representation in State Legislatures." *State Politics & Policy Quarterly* 3 (2): 117–38.

Caress, Stanley. 1999. "The Influence of Term Limits on the Electoral Success of Women." *Women and Politics* 20 (3): 45–63.

Carroll, Susan J., and Krista Jenkins. 2001a. "Do Term Limits Help Women Get Elected?" *Social Science Quarterly* 82 (1): 197–201.

Carroll, Susan J., and Krista Jenkins. 2001b. "Increasing Diversity or More of the Same? Term Limits and the Representation of Women, Minorities, and Minority Women in State Legislatures." Paper presented at the American Political Science Association's Annual Meeting, San Francisco, CA, August 30–September 2.

Carroll, Susan J., and Krista Jenkins. 2001c. "Unrealized Opportunity? Term Limits and the Representation of Women in State Legislatures." *Women and Politics* 23 (4): 1–30.

Ceaser, James W., and Robert P. Saldin. 2005. "A New Measure of Party Strength." *Political Research Quarterly* 58 (2): 245–56.

Center for American Women and Politics. 2005. "Women of Color in State Legislatures 1990, 1995, 2000, and 2005." Personal Request. Eagleton Institute for Politics, Rutgers, State University of New Jersey, New Brunswick, November 2005.

Center for American Women and Politics. 2010a. "Women in State Legislatures 2010." Eagleton Institute for Politics, Rutgers, State University of New Jersey, New Brunswick. http://www.cawp.rutgers.edu/fast_facts/levels_of_office/documents/stleg.pdf (accessed May 4, 2013).

Center for American Women and Politics. 2010b. "Women of Color in Elective Office 2010." Eagleton Institute for Politics, Rutgers, State University of New Jersey, New Brunswick. http://www.cawp.rutgers.edu/fast_facts/levels_of_office/documents/color.pdf (accessed May 4, 2013).

Collins, Todd, and Laura Moyer. 2008. "Gender, Race, and Intersectionality on the Federal Appellate Bench." *Political Research Quarterly* 61 (2): 219–27.

Crenshaw, Kimberle. 1989. "Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics." *University of Chicago Legal Forum*, 139–67.

Crenshaw, Kimberle. 1991. "Mapping the Margins: Intersectionality, Identity Politics, and Violence against Women of Color." *Stanford Law Review* 43 (6): 1241–99.

Darcy, R., and Charles D. Hadley. 1988. "Black Women in Politics: The Puzzle of Success." *Social Science Quarterly* 69 (3): 629–45.

Darling, Marsha. 1998. "African-American Women in State Elective Office in the South." In *Women and Elective Office: Past, Present, and Future*, eds. Sue Thomas and Clyde Wilcox. New York: Oxford University Press, pp. 150–62.

Elazar, Daniel J. [1984] 1966. *American Federalism: A View from the States.* New York: Thomas Y. Crowell.

Erikson, Robert S., Gerald C. Wright, and John P. McIver. 2007. "Replication Data for: Public Opinion in the States: A Quarter Century of Change and Stability." UNF:3:42A1SVhk3cWA2Ss5az8zjQ== Gerald C. Wright [Distributor] V1 [Version]. http://hdl.handle.net/1902.1/10442 (accessed May 4, 2013).

Fraga, Luis Ricardo, Valerie Martinez-Ebers, Linda Lopez, and Ricardo Ramire. 2005. "Strategic Intersectionality: Gender, Ethnicity, and Political Incorporation." Paper delivered at the Western Political Science Association's Annual Meeting, Oakland, March 17-19.

Fraga, Luis Ricardo, Valerie Martinez-Ebers, Ricardo Ramirez, and Linda Lopez. 2003. "Gender and Ethnicity: The Political Incorporation of Latina and Latino State Legislators." Inequality and Social Policy Seminar, John F. Kennedy School of Government, November 10.

Gay, Claudine, and Katherine Tate. 1998. "Doubly Bound: The Impact of Gender and Race on the Politics of Black Women." *Political Psychology* 19 (1): 169–84.

Hardy-Fanta, Carol, Pei-te Lien, Christine Marie Sierra, and Dianne Pinderhughes. 2007. "A New Look at Paths to Political Office: Moving Women of Color from the Margins to the Center." Paper delivered at the American Political Science Association's Annual Meeting, Chicago, August 30–September 2.

Hardy-Fanta, Carol, Christine Marie Sierra, Pei-te Lien, Dianne Pinderhughes, and Wartyna L. Davis. 2005. "Race, Gender, and Descriptive Representation: An Exploratory View of Multicultural Elected Leadership in the United States." Paper delivered at the American Political Science Association's Annual Meeting, Washington, DC, September 1–5.

Hawkesworth, Mary. 2003. "Congressional Enactments of Race-Gender: Toward a Theory of Raced-Gendered Institutions." *American Political Science Review* 97 (4): 529–50.

Hero, Rodney E. 1998. *Faces of Inequality: Social Diversity in American Politics*. New York: Oxford University Press.

Hill, David B. 1981. "Political Culture and Female Political Representation." *The Journal of Politics* 43:159–68.

Hogan, Robert E. 2001. "The Influences of State and District Conditions on the Representation of Women in U.S. State Legislatures." *American Politics Research* 29:4–24.

Johnson, Charles A. 1976. "Political Culture in American States: Elazar's Formulation Examined." *American Journal of Political Science* 20 (3): 491–509.

Lawless, Jennifer L., and Richard L. Fox. 2005. *It Takes a Candidate: Why Women Don't Run for Office*. New York, NY: Cambridge University Press.

Moncrief, Gary, Joel Thompson, and Robert Schuhmann. 1991. "Gender, Race, and the State Legislature: A Research Note on the Double Disadvantage Hypothesis." *Social Science Journal* 28:481–87.

Nechemias, Carol. 1987. "Changes in the Election of Women to U.S. State Legislative Seats." *Legislative Studies Quarterly* 8 (1): 125–142.

Norrander, Barbara, and Clyde Wilcox. 1998. "The Geography of Gender Power: Women in State Legislatures." In *Women and Elective Office: Past, Present, and Future*, eds. by Sue Thomas and Clyde Wilcox. New York: Oxford University Press, pp. 103–17.

Norrander, Barbara, and Clyde Wilcox. 2005. "Change in Continuity in the Geography of Women State Legislators." In *Women and Elective Office: Past, Present, and Future*. 2nd ed., eds. by Sue Thomas and Clyde Wilcox. New York: Oxford University Press, pp. 176–96.

Palmer, Barbara, and Dennis Simon. 2008. *Breaking the Political Glass Ceiling: Women and Congressional Elections*. 2nd ed. New York: Routledge.

Philpot, Tasha S., and Hanes Walton, Jr. 2007. "One of Our Own: Black Female Candidates and the Voters Who Support Them." *American Journal of Political Science* 51 (1): 49–62.

Prestage, Jewel L. 1991. "In Quest of African American Political Woman." *Annals of the American Academy of Political and Social Science* 515: 88–103.

Prindeville, Diane-Michele, and Teresa Braley Gomez. 1991. "American Indian Women Leaders, Public Policy, and the Importance of Gender and Ethnic Identity." *Women and Politics* 20 (2): 17–32.

Rule, Wilma. 1990. "Why More Women Are State Legislators: A Research Note." *Western Political Quarterly* 43 (2): 432–48.

Rule, Wilma. 1999. "Why Are More Women State Legislators?" In *Women in Politics: Outsiders or Insiders?* 3rd ed., ed. Lois Duke Whitaker. Upper Saddle River, NJ: Prentice Hall, pp. 190–202.

Sanbonmatsu, K. 2002. "Political Parties and the Recruitment of Women to State Legislatures." *Journal of Politics* 64 (3): 791–809.

Sanbonmatsu, K. 2006. *Where Women Run: Gender and Party in the American States*. Ann Arbor: University of Michigan Press.

Sharkansky, Ira. 1969. "The Utility of Elazar's Political Culture: A Research Note." *Polity* 2: 66–83.

Smooth, Wendy. 2001. "African American Women State Legislators: The Impact of Gender and Race on Legislative Influence." Doctoral diss. University of Maryland College Park.

Smooth, Wendy. 2006. "Intersectionality in Electoral Politics: A Mess Worth Making." *Politics & Gender* 2 (3): 400–14.

Squire, Peverill. 1992. "Legislative Professionalization and Membership Diversity in State Legislatures." *Legislative Studies Quarterly* 17(1): 69–79.

Squire, Peverill. 2007. "Measuring State Legislative Professionalism: The Squire Index Revisited." *State Politics and Policy Quarterly* 7(2): 211–27.

Williams, Linda Faye. 2001. "The Civil Rights-Black Power Legacy: Black Women Elected Officials at the Local, State, and National Levels." In *Sisters in the Struggle: African American Women in the Civil Rights-Black Power Movement*, eds. Bettye Collier-Thomas and V. P. Franklin. New York: New York University Press, pp. 306–31.

United States Census Bureau. 2000. "United States Census 2000, Summary File 3." http://www. census.gov/census2000/sumfile3.html (accessed May 4, 2013).

United States Census Bureau. 1990. "United States Census 1990, Summary File 3." http://factfinder.census.gov/servlet/DatasetMainPageServlet?_ds_name=DEC_1990_STF3_&_program=DEC&_lang=en (accessed May 4, 2013).

Author Biography

Becki Scola is an assistant professor of political science at Saint Joseph's University. Her research interests include American institutions, gender politics, and race/ethnic politics, and her work has been published in the *Journal of Women, Politics & Policy*, and *Politics & Gender*.

Appendix A

Normal Curve Tail Probabilities. Standard Normal Probability in Right-Hand Tail (for negative values of *z*, probabilities are found by symmetry)

	Second decimal place of z									
z	**.00**	**.01**	**.02**	**.03**	**.04**	**.05**	**.06**	**.07**	**.08**	**.09**
0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641
0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0722	.0708	.0694	.0681
1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367

(Continued)

(Continued)

Second decimal place of z										
z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
1.8	.0359	.0352	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
2.9	.0019	.0018	.0017	.0017	.0016	.0016	.0015	.0015	.0014	.0014
3.0	.00135									
3.5	.000233									
4.0	.0000317									
4.5	.00000340									
5.0	.000000287									

Source: R. E. Walpole, *Introduction to Statistics* (New York: Macmillan, 1968). Used with permission.

Appendix B

Chi-Squared Distribution Values for Various Right-Tail Probabilities

	Alpha level for one-tailed test						
	.05	.025	.01	.005	.0025	.001	.0005
Degree of freedom (df)	Alpha level for two-tailed test						
	.10	.05	.02	.01	.005	.002	.001
1	6.314	12.706	31.821	63.657	127.32	318.31	636.62
2	2.920	4.303	6.965	9.925	14.089	22.327	31.598
3	2.353	3.182	4.541	5.841	7.453	10.214	12.924
4	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8	1.869	2.306	2.896	3.355	3.833	4.501	5.041
9	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	1.812	2.228	2.764	3.169	3.581	4.144	4.587
11	1.796	2.201	2.718	3.106	3.497	4.025	4.437
12	1.782	2.179	2.681	3.055	3.428	3.930	4.318
13	1.771	2.160	2.650	3.012	3.372	3.852	4.221
14	1.761	2.145	2.624	2.977	3.326	3.787	4.140
15	1.753	2.131	2.602	2.947	3.286	3.733	4.073
16	1.746	2.120	2.583	2.921	3.252	3.686	4.015
17	1.740	2.110	2.567	2.898	3.222	3.646	3.965
18	1.734	2.101	2.552	2.878	3.197	3.610	3.922
19	1.729	2.093	2.539	2.861	3.174	3.579	3.883
20	1.725	2.086	2.528	2.845	3.153	3.552	3.850
21	1.721	2.080	2.518	2.831	3.135	3.527	3.819

(Continued)

(Continued)

	Alpha level for one-tailed test						
	.05	.025	.01	.005	.0025	.001	.0005
Degree of freedom (df)	Alpha level for two-tailed test						
	.10	.05	.02	.01	.005	.002	.001
22	1.717	2.074	2.508	2.819	3.119	3.505	3.792
23	1.714	2.069	2.500	2.807	3.104	3.485	3.767
24	1.711	2.064	2.492	2.797	3.091	3.467	3.745
25	1.708	2.060	2.485	2.787	3.078	3.450	3.725
26	1.706	2.056	2.479	2.779	3.067	3.435	3.707
27	1.703	2.052	2.473	2.771	3.057	3.421	3.690
28	1.701	2.048	2.467	2.763	3.047	3.408	3.674
29	1.699	2.045	2.462	2.756	3.038	3.396	3.659
30	1.697	2.042	2.457	2.750	3.030	3.385	3.646
40	1.684	2.021	2.423	2.704	2.971	3.307	3.551
60	1.671	2.000	2.390	2.660	2.915	3.232	3.460
120	1.658	1.980	2.358	2.617	2.860	3.160	3.373
∞	1.645	1.960	2.326	2.576	2.807	3.090	3.291

Source: James V. Couch, *Fundamentals of Statistics for the Behavioral Sciences*, 2nd ed. (St. Paul, Minn.: West, 1987), 327.

Appendix C

Chi-Squared Distribution Values for Various Right-Tail Probabilities

	Right-tail probability						
df	**0.250**	**0.100**	**0.050**	**0.025**	**0.010**	**0.005**	**0.001**
1	1.32	2.71	3.84	5.02	6.63	7.88	10.83
2	2.77	4.61	5.99	7.38	9.21	10.60	13.82
3	4.11	6.25	7.81	9.35	11.34	12.84	16.27
4	5.39	7.78	9.49	11.14	13.28	14.86	18.47
5	6.63	9.24	11.07	12.83	15.09	16.75	20.52
6	7.84	10.64	12.59	14.45	16.81	18.55	22.46
7	9.04	12.02	14.07	16.01	18.48	20.28	24.32
8	10.22	13.36	15.51	17.53	20.09	21.96	26.12
9	11.39	14.68	16.92	19.02	21.67	23.59	27.88
10	12.55	15.99	18.31	20.48	23.21	25.19	29.59
11	13.70	17.28	19.68	21.92	24.72	26.76	31.26
12	14.85	18.55	21.03	23.34	26.22	28.30	32.91
13	15.98	19.81	22.36	24.74	27.69	29.82	34.53
14	17.12	21.06	23.68	26.12	29.14	31.32	36.12
15	18.25	22.31	25.00	27.49	30.58	32.80	37.70
16	19.37	23.54	26.30	28.85	32.00	34.27	39.25
17	20.49	24.77	27.59	30.19	33.41	35.72	40.79
18	21.60	25.99	28.87	31.53	34.81	37.16	42.31
19	22.72	27.20	30.14	32.85	36.19	38.58	43.82
20	23.83	28.41	31.41	34.17	37.57	40.00	45.32
25	29.34	34.38	37.65	40.65	44.31	46.93	52.62
30	34.80	40.26	43.77	46.98	50.89	53.67	59.70
40	45.62	51.80	55.76	59.34	63.69	66.77	73.40

(Continued)

(Continued)

Right-tail probability							
df	0.250	0.100	0.050	0.025	0.010	0.005	0.001
50	56.33	63.17	67.50	71.42	76.15	79.49	86.66
60	66.98	74.40	79.08	83.30	88.38	91.95	99.61
70	77.58	85.53	90.53	95.02	100.4	104.2	112.3
80	88.13	96.58	101.8	106.6	112.3	116.3	124.8
90	98.65	107.6	113.1	118.1	124.1	128.3	137.2
100	109.1	118.5	124.3	129.6	135.8	140.2	149.5

Source: Alan Agresti and Barbara Finlay, *Statistical Methods for the Social Sciences*, 3rd ed. (Upper Saddle River, N.J.: Prentice Hall, 1997), 670. Used with permission.

Appendix D

F Distribution

a =.05										
	df_1									
df_2	1	2	3	4	5	6	8	12	24	∞
1	161.4	199.5	215.7	224.6	230.2	234.0	238.9	243.9	249.0	254.3
2	18.51	19.00	19.16	19.25	19.30	19.33	19.37	19.41	19.45	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.84	8.74	8.64	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.04	5.91	5.77	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.82	4.68	4.53	4.36
6	5.99	5.14	4.76	4.53	4.39	4.28	4.15	4.00	3.84	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.73	3.57	3.41	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.44	3.28	3.12	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.23	3.07	2.90	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.07	2.91	2.74	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	2.95	2.79	2.61	2.40
12	4.75	3.88	3.49	3.26	3.11	3.00	2.85	2.69	2.50	2.30
13	4.67	3.80	3.41	3.18	3.02	2.92	2.77	2.60	2.42	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.70	2.53	2.35	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.64	2.48	2.29	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.59	2.42	2.24	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.55	2.38	2.19	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.51	2.34	2.15	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.48	2.31	2.11	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.45	2.28	2.08	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.42	2.25	2.05	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.40	2.23	2.03	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.38	2.20	2.00	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.36	2.18	1.98	1.73

(Continued)

(Continued)

a =.05										
	df_1									
df_2	1	2	3	4	5	6	8	12	24	∞
25	4.24	3.38	2.99	2.76	2.60	2.49	2.34	2.16	1.96	1.71
26	4.22	3.37	2.98	2.74	2.59	2.47	2.32	2.15	1.95	1.69
27	4.21	3.35	2.96	2.73	2.57	2.46	2.30	2.13	1.93	1.67
28	4.20	3.34	2.95	2.71	2.56	2.44	2.29	2.12	1.91	1.65
29	4.18	3.33	2.93	2.70	2.54	2.43	2.28	2.10	1.90	1.64
30	4.17	3.32	2.92	2.69	2.53	2.42	2.27	2.09	1.89	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.18	2.00	1.79	1.51
60	4.00	3.15	2.76	2.52	2.37	2.25	2.10	1.92	1.70	1.39
120	3.92	3.07	2.68	2.45	2.29	2.17	2.02	1.83	1.61	1.25
∞	3.84	2.99	2.60	2.37	2.21	2.09	1.94	1.75	1.52	1.00

Source: From Table V of R. A. Fisher and F. Yates, *Statistical Tables for Biological, Agricultural and Medical Research*, published by Longman Group Ltd., London, 1974. Reprinted by permission of Pearson Education Limited.

					a =.01					
					df_1					
df_2	1	2	3	4	5	6	8	12	24	∞
1	4052	4999	5403	5625	5764	5859	5981	6106	6234	6366
2	98.49	99.01	99.17	99.25	99.30	99.33	99.36	99.42	99.46	99.50
3	34.12	30.18	29.46	28.71	28.24	27.91	27.49	27.05	26.60	26.12
4	21.20	18.00	16.69	15.98	15.52	15.21	14.80	14.37	13.93	13.46
5	16.26	13.27	12.06	11.39	10.97	10.67	10.27	9.89	9.47	9.02
6	13.74	10.92	9.78	9.15	8.75	8.47	8.10	7.72	7.31	6.88
7	12.25	9.55	8.45	7.85	7.46	7.19	6.84	6.47	6.07	5.65
8	11.26	8.65	7.59	7.01	6.63	6.37	6.03	5.67	5.28	4.86
9	10.56	8.02	6.99	6.42	6.06	5.80	5.47	5.11	4.73	4.31
10	10.04	7.56	6.55	5.99	5.64	5.39	5.06	4.71	4.33	3.91
11	9.65	7.20	6.22	5.67	5.32	5.07	4.74	4.40	4.02	3.60
12	9.33	6.93	5.95	5.41	5.06	4.82	4.50	4.16	3.78	3.36
13	9.07	6.70	5.74	5.20	4.86	4.62	4.30	3.96	3.59	3.16
14	8.86	6.51	5.56	5.03	4.69	4.46	4.14	3.80	3.43	3.00
15	8.68	6.36	5.42	4.89	4.56	4.32	4.00	3.67	3.29	2.87
16	8.53	6.23	5.29	4.77	4.44	4.20	3.89	3.55	3.18	2.75
17	8.40	6.11	5.18	4.67	4.34	4.10	3.79	3.45	3.08	2.65
18	8.28	6.01	5.09	4.58	4.25	4.01	3.71	3.37	3.00	2.57
19	8.18	5.93	5.01	4.50	4.17	3.94	3.63	3.30	2.92	2.49
20	8.10	5.85	4.94	4.43	4.10	3.87	3.56	3.23	2.86	2.42
21	8.02	5.78	4.87	4.37	4.04	3.81	3.51	3.17	2.80	2.36
22	7.94	5.72	4.82	4.31	3.99	3.76	3.45	3.12	2.75	2.31
23	7.88	5.66	4.76	4.26	3.94	3.71	3.41	3.07	2.70	2.26

(Continued)

(Continued)

	a =.01									
	df_1									
df_2	1	2	3	4	5	6	8	12	24	∞
24	7.82	5.61	4.72	4.22	3.90	3.67	3.36	3.03	2.66	2.21
25	7.77	5.57	4.68	4.18	3.86	3.63	3.32	2.99	2.62	2.17
26	7.72	5.53	4.64	4.14	3.82	3.59	3.29	2.96	2.58	2.13
27	7.68	5.49	4.60	4.11	3.78	3.56	3.26	2.93	2.55	2.10
28	7.64	5.45	4.57	4.07	3.75	3.53	3.23	2.90	2.52	2.06
29	7.60	5.42	4.54	4.04	3.73	3.50	3.20	2.87	2.49	2.03
30	7.56	5.39	4.51	4.02	3.70	3.47	3.17	2.84	2.47	2.01
40	7.31	5.18	4.31	3.83	3.51	3.29	2.99	2.66	2.29	1.80
60	7.08	4.98	4.13	3.65	3.34	3.12	2.82	2.50	2.12	1.60
120	6.85	4.79	3.95	3.48	3.17	2.96	2.66	2.34	1.95	1.38
∞	6.64	4.60	3.78	3.32	3.02	2.80	2.51	2.18	1.79	1.00

a = .001										
	df_1									
df_2	1	2	3	4	5	6	8	12	24	∞
1	405284	500000	540379	562500	576405	585937	598144	610667	623497	636619
2	998.5	999.0	999.2	999.2	999.3	999.3	999.4	999.4	999.5	999.5
3	167.5	148.5	141.1	137.1	134.6	132.8	130.6	128.3	125.9	123.5
4	74.14	61.25	56.18	53.44	51.71	50.53	49.00	47.41	45.77	44.05
5	47.04	36.61	33.20	31.09	29.75	28.84	27.64	26.42	25.14	23.78
6	35.51	27.00	23.70	21.90	20.81	20.03	19.03	17.99	16.89	15.75
7	29.22	21.69	18.77	17.19	16.21	15.52	14.63	13.71	12.73	11.69
8	25.42	18.49	15.83	14.39	13.49	12.86	12.04	11.19	10.30	9.34
9	22.86	16.39	13.90	12.56	11.71	11.13	10.37	9.57	8.72	7.81
10	21.04	14.91	12.55	11.28	10.48	9.92	9.20	8.45	7.64	6.76
11	19.69	13.81	11.56	10.35	9.58	9.05	8.35	7.63	6.85	6.00
12	18.64	12.97	10.80	9.63	8.89	8.38	7.71	7.00	6.25	5.42
13	17.81	12.31	10.21	9.07	8.35	7.86	7.21	6.52	5.78	4.97
14	17.14	11.78	9.73	8.62	7.92	7.43	6.80	6.13	5.41	4.60
15	16.59	11.34	9.34	8.25	7.57	7.09	6.47	5.81	5.10	4.31
16	16.12	10.97	9.00	7.94	7.27	6.81	6.19	5.55	4.85	4.06
17	15.72	10.66	8.73	7.68	7.02	6.56	5.96	5.32	4.63	3.85
18	15.38	10.39	8.49	7.46	6.81	6.35	5.76	5.13	4.45	3.67
19	15.08	10.16	8.28	7.26	6.61	6.18	5.59	4.97	4.29	3.52
20	14.82	9.95	8.10	7.10	6.46	6.02	5.44	4.82	4.15	3.38
21	14.59	9.77	7.94	6.95	6.32	5.88	5.31	4.70	4.03	3.26
22	14.38	9.61	7.80	6.81	6.19	5.76	5.19	4.58	3.92	3.15
23	14.19	9.47	7.67	6.69	6.08	5.65	5.09	4.48	3.82	3.05
24	14.03	9.34	7.55	6.59	5.98	5.55	4.99	4.39	3.74	2.97

(Continued)

(Continued)

	a = .001									
	df$_1$									
df$_2$	**1**	**2**	**3**	**4**	**5**	**6**	**8**	**12**	**24**	**∞**
25	13.88	9.22	7.45	6.49	5.88	5.46	4.91	4.31	3.66	2.89
26	13.74	9.12	7.36	6.41	5.80	5.38	4.83	4.24	3.59	2.82
27	13.61	9.02	7.27	6.33	5.73	5.31	4.76	4.17	3.52	2.75
28	13.50	8.93	7.19	6.25	5.66	5.24	4.69	4.11	3.46	2.70
29	13.39	8.85	7.12	6.19	5.59	5.18	4.64	4.05	3.41	2.64
30	13.29	8.77	7.05	6.12	5.53	5.12	4.58	4.00	3.36	2.59
40	12.61	8.25	6.60	5.70	5.13	4.73	4.21	3.64	3.01	2.23
60	11.97	7.76	6.17	5.31	4.76	4.37	3.87	3.31	2.69	1.90
120	11.38	7.31	5.79	4.95	4.42	4.04	3.55	3.02	2.40	1.56
∞	10.83	6.91	5.42	4.62	4.10	3.74	3.27	2.74	2.13	1.00

GLOSSARY

Accretion measures. Measures of phenomena through observation of the accumulation of materials.

Actions. Human behavior done for a reason.

Alternative-form method. A method of calculating reliability by repeating different but equivalent measures at two or more points in time.

Alternative hypothesis. A statement about the value or values of a population parameter. A hypothesis proposed as an alternative to the null hypothesis.

Analysis of variance (ANOVA). A technique for measuring the relationship between one nominal- or ordinal-level variable and one interval- or ratio-level variable.

Antecedent variable. An independent variable that precedes other independent variables in time.

Applied research. Research designed to produce knowledge useful in altering a real-world condition or situation.

Arrow diagram. A pictorial representation of a researcher's explanatory scheme.

Bar graph. A graphic display of the data in a frequency or percentage distribution.

Bias. A type of measurement error that results in systematically over- or under-measuring the value of a concept.

Branching question. A question that sorts respondents into subgroups and directs these subgroups to different parts of the questionnaire.

Case study design. A comprehensive and in-depth study of a single case or several cases. A nonexperimental design in which the investigator has little control over events.

Central tendency. The most frequent, middle, or central value in a frequency distribution.

Chi square. A statistic used to test whether a relationship is statistically significant in a cross-tabulation table.

Classical randomized experimental design. An experiment with the random assignment of subjects to experimental and control groups with a pretest and posttest for both groups.

Closed-ended question. A question with response alternatives provided.

Cluster sample. A probability sample that is used when no list of elements exists. The sampling frame initially consists of clusters of elements.

Cohort. A group of people who all experience a significant event in roughly the same time frame.

Confidence interval. The range of values into which a population parameter is likely to fall for a given level of confidence.

Confidence level. The degree of belief or probability that an estimated range of values includes or covers the population parameter.

Constant. A concept or variable whose values do not vary.

Construct validity. Validity demonstrated for a measure by showing that it is related to the measure of another concept.

Constructionism. An approach to knowledge that asserts humans actually construct—through their social interactions and cultural and historical practices—many of the facts they take for granted as having an independent, objective, or material reality.

Content analysis. A systematic procedure by which records are transformed into quantitative data.

Content validity. Validity demonstrated by ensuring that the full domain of a concept is measured.

Control by grouping. A form of statistical control in which observations identical or similar to the control variable are grouped together.

Control group. A group of subjects that does not receive the experimental treatment or test stimulus.

Convenience sample. A nonprobability sample in which the selection of elements is determined by the researcher's convenience.

Convergent construct validity. Validity demonstrated by showing that the measure of a concept is related to the measure of another, related concept.

Correlation. A statement that the values or states of one thing systematically vary with the values or states of another; an association between two variables.

Correlation coefficient. In regression analysis, a measure of the strength and direction of the linear correlation between two quantitative variables; also called product-moment correlation, Pearson's *r*, or *r*.

Correlation matrix. A table showing the relationships among discrete measures.

Covert observation. Observation in which the observer's presence or purpose is kept secret from those being observed.

Critical theory. The philosophical stance that disciplines such as political science should assess society critically and seek to improve it, not merely study it objectively.

Cross-level analysis. The use of data at one level of aggregation to make inferences at another level of aggregation.

Cross-sectional design. A research design in which measurements of independent and dependent variables are taken at the same time; naturally occurring differences in the independent variable are used to create quasi-experimental and quasi-control groups; extraneous factors are controlled for by statistical means.

Cross-tabulation. Also called a cross-classification or contingency table, this array displays the joint frequencies and relative frequencies of two categorical (nominal or ordinal) variables.

Cumulative. Characteristic of scientific knowledge; new substantive findings and research techniques are built upon those of previous studies.

Cumulative proportion. The total proportion of observations at or below a value in a frequency distribution.

Data matrix. An array of rows and columns that stores the values of a set of variables for all the cases in a dataset.

Deduction. A process of reasoning from a theory to specific observations.

Degrees of freedom. A measure used in conjunction with chi square and other measures to determine if a relationship is statistically significant.

Demand characteristics. Aspects of the research situation that cause participants to guess the purpose or rationale of the study and adjust their behavior or opinions accordingly.

Dependent variable. The phenomenon thought to be influenced, affected, or caused by some other phenomenon.

Descriptive statistic. A number that, because of its definition and formula, describes certain characteristics or properties of a batch of numbers.

Dichotomous variable. A nominal-level variable having only two categories that for certain analytical purposes can be treated as a quantitative variable.

Difference-of-means test. A technique for measuring the relationship between one nominal- or ordinal-level variable and one interval- or ratio-level variable.

Direct observation. Actual observation of behavior.

Direction of a relationship. An indication of which values of the dependent variable are associated with which values of the independent variable.

Directional hypothesis. A hypothesis that specifies the expected relationship between two or more variables.

Discriminant construct validity. Validity demonstrated by showing that the measure of a concept has a low correlation with the measure of another concept that is thought to be unrelated.

Dispersion. The distribution of data values around the most frequent, middle, or central value.

Disproportionate sample. A stratified sample in which elements sharing a characteristic are underrepresented or overrepresented in the sample.

Double-barreled question. A question that is really two questions in one.

Dummy variable. A hypothetical index that has just two values: 0 for the presence (or absence) of a factor and 1 for its absence (or presence).

Ecological fallacy. The fallacy of deducing a false relationship between the attributes or behavior of individuals based on observing that relationship for groups to which the individuals belong.

Ecological inference. The process of inferring a relationship between characteristics of individuals based on group or aggregate data.

Effect size. How and how much a change in one variable affects another variable, often measured as the difference between one mean and another, often between a treatment group and control group.

Electronic databases. A collection of information (of any type) stored on an electromagnetic medium that can be accessed and examined by certain computer programs.

Element. A particular case or entity about which information is collected; the unit of analysis.

Empirical frequency distribution (*f*). The number of observations per value or category of a variable.

Empirical research. Research based on actual, "objective" observation of phenomena.

Empiricism. Relying on observation to verify propositions.

Episodic record. Record that is not part of a regular, ongoing record-keeping enterprise but instead is produced and preserved in a more casual, personal, or accidental manner.

Erosion measures. Measures of phenomena through indirect observation of selective wear of some material.

Estimator. A statistic based on sample observations that is used to estimate the numerical value of an unknown population parameter.

Eta-squared. A measure of association used with the analysis of variance that indicates what proportion of the variance in the dependent variable is explained by the variance in the independent variable.

Ethnography. A type of field study in which the researcher is deeply immersed in the place and lives of the people being studied.

Expected value. The mean or average value of a sample statistic based on repeated samples from a population.

Experiment. Research using a research design in which the researcher controls exposure to the test factor or independent variable, the assignment of subjects to groups, and the measurement of responses.

Experimental effect. Effect, usually measured numerically, of the experimental variable on the dependent variable.

Experimental group. A group of subjects that receives the experimental treatment or test stimulus.

Experimental mortality. A differential loss of subjects from experimental and control groups that affects the equivalency of groups; threat to internal validity.

Explanatory. Characteristic of scientific knowledge; signifying that a conclusion can be derived from a set of general propositions and specific initial considerations; providing a systematic, empirically verified understanding of why a phenomenon occurs as it does.

External validity. The ability to generalize from one set of research findings to other situations.

Face validity. Validity asserted by arguing that a measure corresponds closely to the concept it is designed to measure.

Factor analysis. A statistical technique useful in the construction of multi-item scales to measure abstract concepts.

Falsifiability. A property of a statement or hypothesis such that it can (in principle, at least) be rejected in the face of contravening evidence.

Field experiment. Experimental designs applied in a natural setting.

Field study. Open-ended and wide-ranging (rather than structured) observation in a natural setting.

Filter question. A question used to screen respondents so that subsequent questions will be asked only of certain respondents for whom the questions are appropriate.

Focused interview. A semistructured or flexible interview schedule used when interviewing elites.

Goodman and Kruskal's gamma. A measure of association between ordinal-level variables.

Goodman and Kruskal's lambda. A measure of association between one nominal- or ordinal-level variable and one nominal-level variable.

Guttman scale. A multi-item measure in which respondents are presented with increasingly difficult measures of approval for an attitude.

Histogram. A type of bar graph in which the height and area of the bars are proportional to the frequencies in each category of a categorical variable or intervals of a continuous variable.

Hypothesis. A tentative or provisional or unconfirmed statement that can (in principle) be verified.

Independent variable. The phenomenon thought to influence, affect, or cause some other phenomenon.

Indirect observation. Observation of physical traces of behavior.

Induction. A process of reasoning in which one draws an inference from a set of premises and observations; the premises of an inductive argument support its conclusion but do not prove it.

Informants. Persons who are willing to be interviewed about the activities and behavior of themselves and of the group to which they belong. An informant also helps the researcher engaged in participant observation to interpret group behavior.

Informed consent. Procedures that inform potential research subjects about the proposed research in which they are being asked to participate; the principle that researchers must obtain the freely given consent of human subjects before they participate in a research project.

Institutional review board. Panel to which researchers must submit descriptions of proposed research involving human subjects for the purpose of ethics review.

Interaction. The strength and direction of a relationship depend on an additional variable or variables.

Intercoder reliability. Demonstration that multiple analysts, following the same content analysis procedure, agree and obtain the same measurements.

Interitem association. A test of the extent to which the scores of several items, each thought to measure the same concept, are the same. Results are displayed in a correlation matrix.

Internal validity. The ability to show that manipulation or variation of the independent variable actually causes the dependent variable to change.

Interpretation. Philosophical approach to the study of human behavior that claims that one must understand the way individuals see their world in order to truly understand their behavior or actions; philosophical objection to the empirical approach to political science.

Interval measurement. A measure for which a one-unit difference in scores is the same throughout the range of the measure.

Intervening variable. A variable coming between an independent variable and a dependent variable in an explanatory scheme.

Intervention analysis. A nonexperimental time series design in which measurements of a dependent variable are taken both before and after the "introduction" of an independent variable.

Interviewer bias. The interviewer's influence on the respondent's answers; an example of reactivity.

Interviewing. Interviewing respondents in a nonstandardized, individualized manner.

Kendall's tau-*b* and tau-*c*. Measures of association between ordinal-level variables.

Leading question. A question that encourages the respondent to choose a particular response.

Level of measurement. The extent or degree to which the values of variables can be compared and mathematically manipulated.

Likert scale. A multi-item measure in which the items are selected based on their ability to discriminate between those scoring high and those scoring low on the measure.

Literature review. A systematic examination and interpretation of the literature for the purpose of informing further work on a topic.

Logistic regression. A nonlinear regression model that relates a set of explanatory variables to a dichotomous dependent variable.

Logistic regression coefficient. A multiple regression coefficient based on the logistic model.

Mean. The sum of the values of a variable divided by the number of values.

Measurement. The process by which phenomena are observed systematically and represented by scores or numerals.

Measures of association. Statistics that summarize the relationship between two variables.

Median. The category or value above and below which one-half of the observations lie.

Mode. The category with the greatest frequency of observations.

Mokken scale. A type of scaling procedure that assesses the extent to which there is order in the responses of respondents to multiple items. Similar to Guttman scaling.

Multiple-group design. Experimental design with more than one control and experimental group.

Multiple regression analysis. A technique for measuring the mathematical relationships between more than one independent variable and a dependent variable while controlling for all other independent variables in the equation.

Multiple regression coefficient. A number that tells how much *Y* will change for a one-unit change in a particular independent variable, if all of the other variables in the model have been held constant.

Multivariate analysis. Data analysis techniques designed to test hypotheses involving more than two variables.

Multivariate cross-tabulation. A procedure by which cross-tabulation is used to control for a third variable.

Natural experiment. A study in which comparisons are made among "naturally" occurring groups on variables that cannot be controlled by the investigator.

Negative relationship. A relationship in which high values of one variable are associated with low values of another variable.

Negatively skewed. A distribution of values in which fewer observations lie to the left of the middle value, and those observations are fairly distant from the mean.

Nominal measurement. A measure for which different scores represent different, but not ordered, categories.

Nonnormative knowledge. Knowledge concerned not with evaluation or prescription but with factual or objective determinations.

Nonprobability sample. A sample for which each element in the total population has an unknown probability of being selected.

Normal distribution. A distribution defined by a mathematical formula and the graph of which has a symmetrical bell shape in which the mean, the mode, and the median coincide, and in which a fixed proportion of observations lies between the mean and any distance from the mean measured in terms of the standard deviation.

Normative knowledge. Knowledge that is evaluative, value-laden, and concerned with prescribing what ought to be.

Null hypothesis. A statement that a population parameter equals a single or specific value; the hypothesis that there is no relationship between two variables in the target population. Often, a statement that the difference between two populations is zero.

Open-ended question. A question with no response alternatives provided for the respondent.

Operational definition. The rules by which a concept is measured and scores assigned.

Operationalization. The process of assigning numerals or scores to a variable to represent the values of a concept.

Ordinal measurement. A measure for which the scores represent ordered categories that are not necessarily equidistant from each other.

Overt observation. Observation in which those being observed are informed of the observer's presence and purpose.

Parsimony. The principle that among explanations or theories with equal degrees of confirmation, the simplest—the one based on the fewest assumptions and explanatory factors—is to be preferred; sometimes known as Ockham's razor.

Partial regression coefficient. A number that indicates how much a dependent variable would change if an independent variable changed one unit and all other variables in the equation or model were held constant.

Participant observation. Observation in which the observer becomes a regular participant in the activities of those being observed.

Period effect. An indicator or measure of history effects on a dependent variable during a specified time.

Phi. An association measure that adjusts an observed chi-square statistic by the sample size.

Policy evaluation. Objective analysis of economic, political, cultural, or social effects of public policies.

Population. All the cases or observations covered by a hypothesis; all the units of analysis to which a hypothesis applies.

Population parameter. A characteristic or an attribute in a population (not a sample) that can be quantified.

Positive relationship. A relationship in which the values of one variable increase (or decrease) as the values of another variable increase (or decrease); a relationship in which high values of one variable are associated with high values of another variable.

Positively skewed. A distribution of values in which fewer observations lie to the right of the middle value and those observations are fairly distant from the mean.

Posttest design. Research design in which the dependent variable is measured after, but not before, manipulation of the independent variable.

Pretest. Measurement of the dependent variable prior to the administration of the experimental treatment or manipulation of the independent variable.

Primary data. Data recorded and used by the researcher who is making the observations.

Probabilistic explanation. An explanation that does not explain or predict events with 100 percent accuracy.

Probability sample. A sample for which each element in the total population has a known probability of being selected.

Proportionate reduction in error (*PRE*) measure. A measure of association that indicates how much knowledge of the value of the independent variable of a case improves prediction of the dependent variable compared to the prediction of the dependent variable based on no knowledge of the case's value on the independent variable. Examples are Goodman and Kruskal's lambda, Goodman and Kruskal's gamma, eta-squared, and *R*-squared.

Proportionate sample. A probability sample that draws elements from a stratified population at a rate proportional to size of the samples.

Pure, theoretical, or recreational research. Research designed to satisfy one's intellectual curiosity about some phenomenon.

Purposive sample. A nonprobability sample in which a researcher uses discretion in selecting elements for observation.

Push poll. A poll intended not to collect information but to feed respondents (often) false and damaging information about a candidate or cause.

Quasi-experimental design. A research design that includes treatment and control groups to which individuals are not assigned randomly.

Questionnaire design. The physical layout and packaging of a questionnaire.

Question-order effect. The effect on responses of question placement within a questionnaire.

Quota sample. A nonprobability sample in which elements are sampled in proportion to their representation in the population.

Random digit dialing. A procedure used to improve the representativeness of telephone samples by giving both listed and unlisted numbers a chance of selection.

Random measurement error. An error in measurement that has no systematic direction or cause.

Randomization. The random assignment of subjects to experimental and control groups.

Randomized response technique. A method of obtaining accurate answers to sensitive questions that protects the respondent's privacy.

Range. The distance between the highest and lowest values or the range of categories into which observations fall.

Ratio measurement. A measure for which the scores possess the full mathematical properties of the numbers assigned.

Reactivity. Effect of data collection or measurement on the phenomenon being measured.

Regression analysis. A technique for measuring the relationship between two interval- or ratio-level variables.

Regression coefficient. A statistic that tells how much the dependent variable changes per unit change in the independent variable.

Regression constant. Value of the dependent variable when all of the values of the independent variables in the equation equal zero.

Relationship. The association, dependence, or covariance of the values of one variable with the values of another variable.

Relative frequency. Percentage or proportion of total number of observations in a frequency distribution that have a particular value.

Reliability. The extent to which a measure yields the same results on repeated trials.

Repeated-measurement design. A plan that calls for making more than one measure or observation on a dependent variable at different times over the course of the study.

Research design. A plan specifying how the researcher intends to fulfill the goals of the study; a logical plan for testing hypotheses.

Research or alternative hypothesis. The hypothesis that researchers usually hope to reject the null hypothesis in favor of, represented by HA.

Resistant measure. A measure of central tendency that is not sensitive to one or a few extreme values in a distribution.

Response quality. The extent to which responses provide accurate and complete information.

Response rate. The proportion of respondents selected for participation in a survey who actually participate.

Response set. The pattern of responding to a series of questions in a similar fashion without careful reading of each question.

***R*-squared.** The proportion of the total variation in a dependent variable explained by an independent variable.

Running record. A written record that is enduring and easily accessed and covers an extensive period.

Sample. A subset of observations or cases drawn from a specified population.

Sample bias. The bias that occurs whenever some elements of a population are systematically excluded from a sample. It is usually due to an incomplete sampling frame or a nonprobability method of selecting elements.

Sample statistic. The estimator of a population characteristic or attribute that is calculated from sample data.

Sample-population congruence. The degree to which sample subjects represent the population from which they are drawn.

Sampling distribution. A theoretical (nonobserved) distribution of sample statistics calculated on samples of size *N* that, if known, permits the calculation of confidence intervals and the test of statistical hypotheses.

Sampling error. The difference between a sample estimate and a corresponding population parameter that arises because only a portion of a population is observed.

Sampling fraction. The proportion of the population included in a sample.

Sampling frame. The population from which a sample is drawn. Ideally, it is the same as the total population of interest to a study.

Sampling interval. The number of elements in a sampling frame divided by the desired sample size.

Sampling unit. The entity listed in a sampling frame. It may be the same as an element, or it may be a group or cluster of elements.

Scatterplot. A graph that plots joint values of an independent variable along one axis (usually the *x*-axis) and a dependent variable along the other axis (usually the *y*-axis).

Search engine. A computer program that visits Web pages on the Internet and looks for those containing particular directories or words.

Search term. A word or phrase entered into a computer program (a search engine) that looks through Web pages on the Internet for those that contain the word or phrase.

Secondary data. Data used by a researcher that were not personally collected by that researcher.

Selection bias. Bias due to the assignment of subjects to experimental and control groups according to some criterion and not randomly; threat to internal validity.

Simple random sample. A probability sample in which each element has an equal chance of being selected.

Single-sided question. A question in which the respondent is asked to agree or disagree with a single substantive statement.

Small-*N* design. A research design in which the researcher examines one or a few cases of a phenomenon in considerable detail.

Snowball sample. A sample in which respondents are asked to identify additional members of a population.

Social facts. Values and institutions that have a subjective existence in the minds of people living in a particular culture.

Somers' *D*. A measure of association between ordinal-level variables.

Split-halves method. A method of calculating reliability by comparing the results of two equivalent measures made at the same time.

Standard deviation. A measure of dispersion of data points about the mean for interval- and ratio-level data.

Standard error. The standard deviation or measure of variability or dispersion of a sampling distribution.

Standardized regression coefficient. A coefficient that measures the effects of an independent variable on a dependent variable in standard deviation units.

Standardized variable. A rescaled variable obtained by subtracting the mean from each value of the variable and dividing the quotient by the standard deviation.

Statistical hypotheses. Two types of hypotheses essential to hypothesis testing: null hypotheses and research or alternative hypotheses.

Statistical independence. A property of two variables where the probability that an observation is in a particular category of one variable and a particular category of the other variable equals the simple or marginal probability of being in those categories.

Statistical inference. The mathematical theory and techniques for making conjectures about the unknown characteristics (parameters) of populations based on samples.

Statistical significance. The probability of making a type I error.

Stratified sample. A probability sample in which elements sharing one or more characteristics are grouped and elements are selected from each group in proportion to the group's representation in the total population.

Stratum. A subgroup of a population that shares one or more characteristics.

Structured observation. Systematic observation and recording of the incidence of specific behaviors.

Summation index. A multi-item measure in which individual scores on a set of items are combined to form a summary measure.

Survey instrument. The schedule of questions to be asked of the respondent.

Systematic sample. A probability sample in which elements are selected from a list at predetermined intervals.

Tautology. A hypothesis in which the independent and dependent variables are identical, making it impossible to disconfirm.

Test of statistical significance. A convention for testing hypotheses that focuses on the probability of making a type I error.

Test stimulus or test factor. The independent variable introduced and controlled by an investigator in order to assess its effects on a response or dependent variable.

Test-retest method. A method of calculating reliability by repeating the same measure at two or more points in time.

Theory. A statement or series of related statements that organize, explain, and predict phenomena.

Time series design. A research design (sometimes called a longitudinal design) featuring multiple measurements of the dependent variable before and after experimental treatment.

Total variation. A numerical measure of the variation in a variable, determined by summing the squared deviation of each observation from the mean.

Transmissible. Characteristic of scientific knowledge; indicates that the methods used in making scientific discoveries are made explicit so that others can analyze and replicate findings.

Trend analysis. Research design that measures a dependent variable at different times and attempts to determine whether the level of the variable is changing—and, if it is, why.

Two-sided question. A question with two substantive alternatives provided for the respondent.

Type I error. Error made by rejecting a null hypothesis when it is true.

Type II error. Error made by failing to reject a null hypothesis when it is not true.

Unit of analysis. The type of actor (individual, group, institution, nation) specified in a researcher's hypothesis.

Unstructured observation. Observation in which all behavior and activities are recorded.

Validity. The correspondence between a measure and the concept it is supposed to measure.

Variance. A measure of dispersion of data points about the mean for interval- and ratio-level data.

Verification. The process of confirming or establishing a statement with evidence.

Weighting factor. A mathematical factor used to make a disproportionate sample representative.

Written record. Documents, reports, statistics, manuscripts, and other recorded materials available and useful for empirical research.

***z* score.** The number of standard deviations by which a score deviates from the mean score.

INDEX