Gary King: ***A Solution to the Ecological Inference Problem: Reconstructing Individual Behavior from Aggregate Data***
is published by Princeton University Press and copyrighted,
© 1997, Princeton University Press. All rights reserved. This
text may be used and shared in accordance with the fair-use provisions
of US copyright law, and it may be archived and redistributed in
electronic form, provided that this notice is carried, Princeton
University Press is notified, the entire original is distributed
without modification, and no fee is charged for access. Archiving,
redistribution, or republication of this text on other terms, in any
medium, requires the consent of Princeton University Press.

For COURSE PACK PERMISSIONS, refer to entry on previous menu.
For more information, send e-mail to permissions@pupress.princeton.edu

# A Solution to the Ecological Inference Problem

# A Solution to the Ecological Inference Problem

## RECONSTRUCTING INDIVIDUAL BEHAVIOR FROM AGGREGATE DATA

*Gary King*

PRINCETON UNIVERSITY PRESS
PRINCETON, NEW JERSEY

Copyright © 1997 by Princeton University Press
Published by Princeton University Press, 41 William Street,
Princeton, New Jersey 08540
In the United Kingdom: Princeton University Press,
Chichester, West Sussex

All Rights Reserved

*Library of Congress Cataloging-in-Publication Data*

King, Gary.
A solution to the ecological inference problem: reconstructing individual behavior from aggregate data / Gary King.
p. cm.
Includes bibliographical references and index.
ISBN 0-691-01241-5 (alk. paper). — ISBN 0-691-01240-7 (pbk.: alk. paper)
1. Political science—Statistical methods. 2. Inference.
I. Title.
JA71.7.K55 1997
320′.072—dc20 9632986
CIP

This book has been composed in Palatino

Princeton University Press books are printed on
acid-free paper and meet the guidelines
for permanence and durability of the Committee
on Production Guidelines for Book Longevity
of the Council on Library Resources

Printed in the United States of America
by Princeton Academic Press

1 3 5 7 9 10 8 6 4 2
1 3 5 7 9 10 8 6 4 2
(Pbk.)

*For Ella Michelle King*

# Contents

# Figures

# Tables

# Preface

IN THIS BOOK, I present a solution to the ecological inference problem: a method of inferring individual behavior from aggregate data that works in practice. *Ecological inference* is the process of using aggregate (i.e., "ecological") data to infer discrete individual-level relationships of interest when individual-level data are not available. Existing methods of ecological inference generate very inaccurate conclusions about the empirical world—which thus gives rise to the ecological inference *problem*. Most scholars who analyze aggregate data routinely encounter some form of the this problem.

The ecological inference problem has been among the longest standing, hitherto unsolved problems in quantitative social science. It was originally raised over seventy-five years ago as the first statistical problem in the nascent discipline of political science, and it has held back research agendas in most of its empirical subfields. Ecological inferences are required in political science research when individual-level surveys are unavailable (for example, local or comparative electoral politics), unreliable (racial politics), insufficient (political geography), or infeasible (political history). They are also required in numerous areas of major significance in public policy (for example, for applying the Voting Rights Act) and other academic disciplines, ranging from epidemiology and marketing to sociology and quantitative history.[1]

Because the ecological inference problem is caused by the lack of individual-level information, no method of ecological inference, including that introduced in this book, will produce precisely accurate results in every instance. However, potential difficulties are minimized here by models that include more available information, diagnostics to evaluate when assumptions need to be modified, and realistic uncertainty estimates for all quantities of interest. For political methodologists, many opportunities remain, and I hope the

[1] What is "ecological" about the aggregate data from which individual behavior is to be inferred? The name has been used at least since the late 1800s and stems from the word ecology, the science of the interrelationship of living things and their environments. Statistical measures taken at the level of the environment, such as summaries of geographic areas or other aggregate units, are widely known as ecological data. Ecological inference is the process of using ecological data to learn about the behavior of individuals within these aggregates.

results reported here lead to continued research into and further improvements in the methods of ecological inference. But most importantly, the solution to the ecological inference problem presented here is designed so that empirical researchers can investigate substantive questions that have heretofore proved intractable. Perhaps it will also lead to new theories and empirical research in areas where analysts have feared to tread due to the lack of reliable ecological methods or individual-level data.

## OUTLINE

This book is divided into five main parts. Part I contains two introductions. Chapter 1 is a qualitative introduction to the entire book and includes a summary of results, an overview of some of the uses to which the method can be put, and a brief outline of the statistical model; because it includes no technical details about the statistical method developed in the subsequent fifteen chapters, it should be accessible even to those without a background in statistics. Chapter 2 gives a formal statement of the ecological inference problem along with the mathematical notation used throughout the remainder of the book.

Part II is divided into aggregation problems (Chapter 3) and problems unrelated to aggregation (Chapter 4). In the first of these chapters, I prove that all of the diverse problems attributed to aggregation bias in the literature are mathematically equivalent, so that only one aggregation problem remains to be solved. The second of these chapters describes a series of basic statistical problems that, although unrelated to aggregation and mostly ignored in the literature, still affect ecological inferences. Any model intended to provide valid ecological inferences must resolve all non-aggregation problems as well.

Part III describes my proposed solution to the ecological inference problem. It reformulates the data by generalizing the method of bounds both algebraically and with easy-to-use graphical methods as well as providing narrower, more informative bounds for the aggregate-level quantities of interest than have been used in the literature (Chapter 5), and introduces a statistical approach to modeling the remaining uncertainty within the observation-level deterministic bounds (Chapter 6). Chapter 7 develops procedures for estimating the model; Chapter 8 shows how to compute quantities of interest at the aggregate level and for each individual observation. Chapter 9 explains how to verify model assumptions with only aggregate data, shows what can go wrong, and provides diagnostic tests, extensions

of the basic model, and a fully nonparametric version to cope with any problems that may remain. Part III also explains how the ecological inference problem and the solution proposed are mathematically equivalent to aspects of the "tomography" problem, which involves reconstructing cross-sectional images of body parts using X-rays and CT scans rather than surgery, or images of the earth's interior via inferences from the detection of seismic waves, produced by earthquakes or nuclear explosions, instead of by digging.

Part IV validates the model by comparing myriad observation-level estimates from the model using aggregate data with their corresponding, known individual-level truths. These comparisons include a typical example of ecological inference, a study of registration by race in the 1960s Southern United States with all the intermediate results described (Chapter 10); an analysis of poverty status by sex in South Carolina which demonstrates that the model is highly robust to aggregation bias and restricted aggregate variances (Chapter 11); a study of black registration in Kentucky that shows how the model gives reasonable answers even in the face of ecological data with almost all relevant information aggregated away (Chapter 12); and two classic applications of ecological inference, the transitions of voters between elections and turn-of-the-century county data on literacy by race (Chapter 13). The method works in practice: it gives accurate answers and correct assessments of uncertainty even when existing methods lead to incorrect inferences or impossible results (such as $-20\%$ of African Americans voting).

Finally, Part V generalizes the basic model in several ways and then concludes. Chapter 14 analyzes three related non-ecological aggregation problems: solving the "modifiable areal unit problem" (a related problem in geography); combining survey and aggregate data to improve ecological inferences (as often studied in the discipline of statistics); and using aggregate-level data for inferences about relationships among continuous individual-level variables (a standard aggregation problem in econometrics). Chapter 15 generalizes the basic model to larger and multidimensional tables.

Chapter 16 concludes with a checklist of items to consider in applying the methods developed here. Technical appendices and a Glossary of Symbols follow.

## ROADMAPS

This book is intended to be read sequentially, as each chapter builds on the material that comes before. For example, Part I should be

read by all, since it includes an overview, a formal statement of the ecological inference problem, and the notation used throughout the book. Nevertheless, for a first reading, some readers may wish to skip certain passages by following one of the roadmaps provided here.

Although Part II introduces several new results and provides motivation for many features of the solution offered in this book, readers uninterested in prior approaches to ecological inference may wish to skim this part by reading only pages 37–43 and Section 3.5 (on pages 54–55) in Chapter 3, along with the indented, italicized remarks in Chapter 4.

Those readers who wish a quicker introduction to the proposed methods should read Part I, and skim Part II as described above. Then, a brief summary of the most basic form of the statistical model requires the information about the data and bounds in Chapter 5 (especially the explanation of Figure 5.1), and the introduction to the model and interpretation on pages 91–96 in Chapter 6. See also Chapter 10 for an application.

All readers should be aware that the solution to the ecological inference problem put forth and verified in this book is more than the basic statistical model that lies at its core. It also includes various extensions to avoid specific problems, a variety of new diagnostic procedures, graphical techniques, and methods of interpretation. Each of these, discussed in the rest of the book, is an integral part of making valid inferences about relationships among individual variables using only aggregate data. Many of these features of the methodology are demonstrated during the verification of the method in Part IV. Especially important is Chapter 16, which provides a checklist for those who intend to use these methods.

## BACKGROUND

Although I hope the results reported here are useful to technically sophisticated political methodologists in building better models of ecological inference, my primary intended audience for this book is political scientists and others who need to make ecological inferences in real academic research, scholars for whom the substantive answer matters. Thus, the qualitative overview in Chapter 1 assumes no statistical knowledge. Parts I, II, and IV assume familiarity with linear regression. Although Parts III and V introduce a variety of tools to solve the ecological inference problem, most of the exposition assumes knowledge of only basic maximum likelihood methods (such as Cramer, 1986 or King, 1989a).

## SOFTWARE AND DATA

Two versions of an easy-to-use, public-domain computer program that implement all the statistical and graphical methods proposed herein are available from my homepage on the World Wide Web at `http://GKing.Harvard.Edu`. One version, *EI: A Program for Ecological Inference*, works under the Gauss software package and is also distributed with Gauss, as part of its Constrained Maximum Likelihood module.[2] The other version, *$E_z$I: A(n easy) Program for Ecological Inference*, by Ken Benoit and me, is a stand-alone, menu-based system that is less flexible but does not require any other software. The methods introduced here are also being incorporated in several general-purpose statistical packages; when these are complete, I will list this information at my homepage.

In order to meet the replication standard (King, 1995), I have deposited all data used in this manuscript, all computer software written for it, and all additional information necessary to replicate the empirical results herein with Inter-University Consortium for Political and Social Research (ICPSR) in their Publication-Related Archive.

## ACKNOWLEDGMENTS

Like many political methodologists, I have been interested in the ecological inference problem for a long time and have benefited from innumerable conversations with many colleagues. I wrote several papers on the subject while an undergraduate, and my first paper in graduate school and my first paper after graduate school attempted to make inferences from aggregate data to individuals. My thanks to Leon Epstein, Bert Kritzer, and Jerry Benjamin for teaching me about the problems in the literature and the more serious ones in my papers. For most of the years since, I followed the literature while working on other projects, but my interest was heightened by a grant I received at the end of the 1990 redistricting process from a now-defunct nonpartisan foundation. This grant included a donation of the largest set of U.S. precinct-level electoral data ever assembled in one place (originally collected to aid minorities during redistricting). The U.S. National Science Foundation provided another key grant (SBR-9321212) to enable me over the next several years to clean these data, merge them with U.S. Census data, and develop ecological inference methods for their analysis. My thanks goes to officials at the

[2] Gauss is available from Aptech Systems, Inc.; 23804 S.E. Kent-Kangley Road; Maple Valley, Washington 98038; (206) 432-7855; `sales@aptech.com`.

NSF Programs in Political Science (Frank Scioli, John McIver), Geography (J. W. Harrington, David Hodge), and Methods, Models, and Measurement (Cheryl Eavey) for arranging the grant, and for a creative conspiracy to introduce me to geographers working on related problems. I find that new methods are best developed while analyzing real data, and these data have proved invaluable. Thanks to my colleague Brad Palmquist, who joined with me in leading the data project soon after he arrived at Harvard, and to the extraordinarily talented crew of graduate students—Greg Adams, Ken Benoit, Grant Emison, Debbie Javeline, Claudine Gay, Jeff Lewis, Eric Reinhardt, and Steve Voss—and undergraduate students—Sarah Dix, Jim Goldman, Paul Hatch, and Robert Hutter—for their research assistance and many creative contributions.

Nuffield College at Oxford University provided a visiting fellowship and a wonderful place to think about these issues during the summer of 1994; my appreciation goes to Clive Payne for hosting my visit and for many interesting discussions. The John Simon Guggenheim Memorial Foundation deserves special thanks for a fellowship that enabled me to spend an exciting year (1994–1995) devoted full time, almost without interruption, to this project. The departments of Political Science and Geography at the University of Colorado provided a forum in September 1994, so that I could discuss my initial results. I presented the final version of the statistical model at the March 1995 meetings of the Association of American Geographers, and in May I presented this model along with most of the empirical evidence offered here in the political science and statistics seminar sponsored by the Institute for Governmental Studies at the University of California, Berkeley. My sincere thanks to Luc Anselin for very insightful comments at the geography meetings, and to Andrew Gelman for going beyond the call of duty in helping me correct several proofs and work out some computational problems following my Berkeley presentation. My first exposure to tomographic research came from Andrew's dissertation work at Harvard, and my understanding of its use as a heuristic device for portraying the model and developing diagnostic procedures came during our conversations and Andrew's many insightful suggestions, some of which are also acknowledged in the text of the book. Without what I learned from Andrew during our collaborations on other projects over the last decade, this book would probably look very different.

I am indebted to the exceptionally talented scholars who attend the always lively summer meetings of the Political Methodology Group. They have provided many years of encouragement, perceptive feedback, and numerous new ideas. The meeting in July 1995 was no

exception: Chris Achen was my formally designated discussant, and numerous others were very helpful at the meetings and almost continuously thereafter. Mike Alvarez, Neal Beck, and John Freeman provided written suggestions and answered numerous questions about many topics. John also gave me some especially helpful comments as my discussant at the American Political Science Association meetings in August 1995. Neal Beck's remarkable ability to read my manuscript and simultaneously to write trenchant e-mail comments and respond to relentless further inquiries about it kept me very busy and improved the final product immeasurably; I have learned a great deal over the years from our frequent electronic communications. Thanks also to Henry Brady, Nancy Burns, Gary Chamberlain, Charles Franklin, Dave Grether, Jerry Hausman, Ken McCue, Doug Rivers, Jim Stock, Wendy Tam, Søren Thomsen, Waldo Tobler and Chris Winship for helpful conversations, to Larry Bartels, Gary Cox, Dudley Duncan, Mitchell Duneier, David Epstein, Sharon O'Halloran, and Bert Kritzer for insightful written comments, to Sander Greenland for helping me learn about epidemiological research, and to Danny Goldhagen for help with the literature on the Nazi vote.

My colleagues Jim Alt, Mo Fiorina, Brad Palmquist, and Sid Verba were very helpful throughout the process, providing many comments, numerous insightful discussions, and much encouragement. Nothing beats the off-hand comments at the mailboxes in the Department of Government. Brad Palmquist's careful readings of the manuscript, many useful suggestions, and deep knowledge of ecological inference saved me from several blunders and improved the final product substantially. Alex Schuessler and Jonathan Katz provided very insightful comments on painfully early versions of the manuscript. Curt Signorino was far more than my lead research assistant; he provided important, perceptive reactions to many of my ideas in their earliest forms, helped me reason through numerous difficult statistical and mathematical issues, corrected several proofs, and helped me work through a variety of computational issues. My appreciation goes to Chuck D'Antonio, Yi Wang, and William Yi Wei for their computer wizardry.

Alison Alter, Ken Scheve (Harvard); Barry Burden, David Kimball, Chris Zorn (OSU); Jeff Lewis, Jason Wittenberg (MIT); Fang Wang (Cal Tech); and other faculty and students deserve special thanks for letting me experiment on them with alternative drafts and computer programs and for their continual inspiration and suggestions. Most of these political scientists participated with me in a "virtual seminar" held over the 1995–1996 academic year: they helped me improve the manuscript by identifying passages that were unclear, unhelpful, or

untrue, and I tried to return the favor with immediate explanations via e-mail of anything that was holding them up. No doubt I got the better end of this bargain!

When the project was farther along still, I had the great fortune to receive comments on this work from presentations I gave at Michigan State University's Political Institutions and Public Choice Program (February 16, 1996), the University of Iowa (February 29, 1996), the Harvard-MIT Econometrics Workshop (April 4, 1996), Columbia University's Center for Social Science (April 5, 1996), the University of California, Santa Barbara (April 10, 1996), the California Institute of Technology (April 11, 1996), the University of California, Los Angeles (April 12, 1996) and the ICPSR program at the University of Michigan (17 July 1996). I am also grateful to Jim Alt, Steve Voss, and Michael Giles and Kaenan Hertz for graciously providing access to their valuable (separate) data sets on race and registration. I especially appreciate the talented staff at Princeton University Press, including Malcolm Litchfield and Peter Dougherty (editors), Jane Low (production manager), and Margaret Case (copy editor), for their professionalism and dedication.

Elizabeth has my deepest appreciation for everything from love and companionship to help with logic and calculus. The dedication is to our daughter, who learned how to laugh just as I was finishing this book. I am reasonably confident that the two events are unrelated, although I will let you know after she learns to talk!

PART I

# Introduction

Chapter 1 provides a qualitative overview of the entire book. It should be accessible even to readers without statistical background. Chapter 2 gives a formal algebraic statement of the ecological inference problem and sets out the basic notation used throughout the book.

CHAPTER 1

# Qualitative Overview

POLITICAL SCIENTISTS have understood the ecological inference problem at least since William Ogburn and Inez Goltra (1919) introduced it in the very first multivariate statistical analysis of politics published in a political science journal (see Gow, 1985; Bulmer, 1984). In a study of the voting behavior of newly enfranchised women in Oregon, they wrote that "even though the method of voting makes it impossible to count women's votes, one wonders if there is not some indirect method of solving the problem. The height of a waterfall is not measured by dropping a line from the top to the bottom, nor is the distance from the earth to the sun measured by a rod and chain" (p. 414).[1]

Ogburn and Goltra's "indirect" method of estimating women's votes was to correlate the percent of women voting in each precinct in Portland, Oregon, with the percent of people voting "no" in selected referenda in the same precincts. They reasoned that individual women were probably casting ballots against the referenda questions at a higher rate than men "if precincts with large percentages of women voting, vote in larger percentages against a measure than the precincts with small percentages of women voting." But they (correctly) worried that what has come to be known as the ecological inference problem might invalidate their analysis: "It is also theoretically possible to gerrymander the precincts in such a way that there may be a negative correlative even though men and women each distribute their votes 50 to 50 on a given measure" (p. 415). The essence of the ecological inference problem is that the true individual-level relationship could even be the reverse of the observed aggregate correlation if it were the *men* in the heavily female precincts who voted disproportionately against the referenda.

Ogburn and Goltra's data no longer appear to be available, but the problem they raised can be illustrated by this simple hypothetical example reconstructed in part from their verbal descriptions. Consider

[1] In 1919, the possibility of what has since come to be known as the "gender gap" was a central issue for academics and a nontrivial concern for political leaders seeking reelection: Not only were women about to have the vote for the first time nationwide; because women made up slightly over fifty percent of the population, they were about to have *most* of the votes.

two equal-sized precincts voting on Proposition 22, an initiative by the radical "People's Power League" to institute proportional representation in Oregon's Legislative Assembly elections: 40% of voters in precinct 1 are women and 40% of all voters in this precinct oppose the referenda. In precinct 2, 60% of voters are women and 60% of the precinct opposes the referenda. Precinct 2 has more women and is more opposed to the referenda than precinct 1, and so it certainly *seems* that women are opposing the proportional representation reform. Indeed, it could be the case that all women were opposed and all men voted in favor in both precincts, as might have occured if the reform were uniformly seen as a way of ensuring men a place in the legislature even though they formed a (slight) minority in every legislative district. But however intuitive this inference may appear, simple arithmetic indicates that it would be equally consistent with the observed aggregate data for men to have opposed proportional representation at a rate four times higher than that of women.[2] These higher relative rates of individual male opposition would occur, given the same aggregate percentages, if a larger fraction of men in the female-dominated precinct 2 opposed the reform than men in precinct 1, as might happen if precinct 2 was a generally more radical area independent of, or even because of, its gender composition.

But if Ogburn and Goltra were Leif Ericson, William Robinson was Christopher Columbus: for not until Robinson's (1950) article was the problem widely recognized and the quest for a valid method of making ecological inferences begun in earnest.[3] Robinson's article remains one of the most influential works in social science methodology. His (correct) view was that, with the methods available at the time, valid ecological inference was impossible. He warned analysts never to use aggregate data to infer individual relationships, and thus to avoid what has since come to be known as "the ecological fallacy." His work

[2] That is, given these aggregate numbers, a minimum of 0% of females in precinct 1 and 20% in precinct 2 (for an average of 10%) could have opposed the referenda, whereas a maximum of 40% of males in each precinct could have opposed it. Chapter 5 provides easy graphical methods of making calculations like these.

[3] Other early works that recognized the ecological inference problem include Allport (1924), Bernstein (1932), Gehlke and Biehl (1934), Thorndike (1939), Deming and Stephan (1940), and Yule and Kendall (1950). Robinson (1950) cited several of these studies as well as Ogburn and Goltra. Scholars writing even earlier than Ogburn and Goltra (1919) made ecological inferences, even though they did not recognize the problems with doing so. In fact, even the works usually cited as the first statistical works of any kind, which incidentally concerned political topics, included ecological inferences (see Graunt, 1662, and Petty, 1690, 1691). See Achen and Shively (1995) for other details of the history of ecological inference research.

sent two shock waves through the social sciences that are still being felt, causing some scholarly pursuits to end and another to begin.

First, the use of aggregate data by political scientists, quantitative historians, sociologists, and others declined relative to use of other forms of data; scholars began to avoid using aggregate data to address whole classes of important research questions (King, 1990). In many countries and fields of study, this "collapse of aggregate data analysis ... and its replacement by individual survey analysis as the dominant method of quantitative social research" (Achen and Shively, 1995: 5) meant that numerous, often historical and geographical, issues were put aside, and many still remain unanswered. What might have become vibrant fields of scholarship withered. The scholars who continue to work in these fields—such as those in comparative politics attempting to explain who voted for the Nazi party, or political historians studying working-class support for political parties in the antebellum Southern U.S.—do so because of the lack of an alternative to ecological data, but they toil under a cloud of great suspicion. The ecological inference problem hinders substantive work in almost every empirical field of political science, as well as numerous areas of sociology, education, marketing, economics, history, geography, epidemiology, and statistics. For example, historical election statistics have fallen into disuse and studies based on them into at least some disrepute. Classic studies, such as V. O. Key's (1949) *Southern Politics*, have been succeeded by scholarship based mostly on survey research, often to great advantage, but necessarily ignoring much of history, focused as it is on the few recent, mostly national, elections for which surveys are available.

The literature's nearly exclusive focus on national surveys with random interviews of isolated individuals means that the geographic component to social science data is often neglected. Commercial state-level surveys are available, but their quality varies considerably and the results are widely suspect in the academic community. Even if the address of each survey respondent were available, the usual 1,000–2,000 respondents to national surveys are insufficient for learning much about spatial variation except for the grossest geographic patterns, in which a country would be divided into no more than perhaps a dozen broad regions. For example, some National Election Study polls locate respondents within congressional districts, but only about a dozen interviews are conducted in any district, and no sample is taken from most of the congressional districts for any one survey. The General Social Survey makes available no geographic information to researchers unless they sign a separate confidentiality agreement, and even then only the respondent's state of residence is released. Survey

organizations in other countries are even more reticent about releasing local geographic information.

Creative combinations of quantitative and qualitative research are much more difficult when the identity and rich qualitative information about individual communities or respondents cannot be revealed to readers. Indeed, in most cases, respondents' identities are not even known to the data analyst. If "all politics is local," political science is missing much of politics. In contrast, aggregate data are saturated with precise spatial information. For example, the United States can be divided into approximately 190,000 electoral precincts, and detailed aggregate political data are available for each. Only the ecological inference problem stands between the scientific community and this rich source of information.

Whereas the first shock wave from Robinson's article stifled research in many substantive fields, the second energized the social science statistics community to try to solve the problem. One partial measure of the level of effort devoted to solving the ecological inference problem is that Robinson's article has been cited more than eight hundred times.[4] Many other scholars have written on the topic as well, citing those who originally cited Robinson or approaching the problem from different perspectives. At one extreme, the literature includes authors such as Bogue and Bogue (1982), who try, unsuccessfully, to "refute" the ecological fallacy altogether; at the other extreme are fatalists who liken the seventy-five year search for a solution to the ecological inference problem to seeking "alchemists' gold" (Flanigan and Zingale, 1985) or to "a fruitless quest" (Achen and Shively, 1995). These scholars, and numerous others between these extreme positions, have written extensively, and often very fruitfully, on the topic. Successive generations of young scholars and methodologists in the making, having been warned off aggregate data analysis with their teachers' mantra "thou shalt not draw conclusions about individual behavior from aggregate data," come away with the conviction that the ecological inference problem presents an enormous barrier to social science research. This belief has drawn a steady stream of social science methodologists into the search for a solution over the years, myself included.

Numerous important advances have been made in the ecological inference literature, but even the best current methods give incorrect answers a large fraction of the time, and nonsensical answers very

[4] This is a vast underestimate, as it depends on data from the *Social Science Citation Index*, which did not even begin publishing (or counting) until six years after Robinson's article appeared.

frequently (such as 115% of blacks voting for the Democrats or −4% of foreign-born Americans being illiterate). No proposed method has been scientifically validated. Any that have been tried on data sets for which the individual-level relationship of interest is known generally fail to give the right answer. It is a testimony to the difficulty of the problem that no serious attempts have even been made to address a variety of basic statistical issues related to the problem. For example, currently available measures of uncertainty, such as confidence intervals, standard errors, and others, have never been validated and appear to be hopelessly inaccurate. Indeed, for some important approaches, no uncertainty measures have even been proposed.

Unlike the rest of this book, this chapter contains no technical details and should be readable even by those with little or no statistical background. In the remainder of this chapter, I summarize some other applications of ecological inference (Section 1.1), define the problem more precisely by way of a leading example of the failures of the most popular current method (Section 1.2), summarize the nature of the solution offered (Section 1.3), provide some brief empirical evidence that the method works in practice (Section 1.4), and outline the statistical method offered (Section 1.5).

## 1.1 The Necessity of Ecological Inferences

Contrary to the pessimistic claims in the ecological inference literature (since Robinson, 1950), aggregate data are sometimes useful even without inferences about individuals. Studies of incumbency advantage, the political effects of redistricting plans, forecasts of macro-economic conditions, and comparisons of infant mortality rates across nations are just a few of the cases where both questions and data coincide at the aggregate level.[5] Nevertheless, even studies such as these that ask questions about aggregates can usually be improved with valid inferences about the individuals who make up the aggregates. And more importantly, numerous other questions exist for which only valid ecological inferences will do.

Fundamental questions in most empirical subfields of political science require ecological inferences. Researchers in many other fields

[5] There are even several largely independent lines of research that give conditions under which aggregate data is not worse than individual-level data for certain purposes. In political science, see Kramer (1983); in epidemiology, see Morgenstern (1982); in psychology, see Epstein (1986); in economics, see Grunfeld and Griliches (1960), Fromm and Schink (1973), Aigner and Goldfeld (1974), and Shin (1987); and in input-output analysis, a field within economics, see Malinvaud (1955) and Venezia (1978).

of academic inquiry, as well as the real world of public policy, also routinely try to make inferences about the attributes of individual behavior from aggregate data. If a valid method of making such inferences were available, scholars could provide accurate answers to these questions with ecological data, and policymakers could base their decisions on reliable scientific techniques. Many of the ecological inferences pursued in these other fields are also of interest to political scientists, which reemphasizes the close historical connection between the ecological inference problem and political science research. The following list represents a small sample of ecological inferences that have been attempted in a variety of fields.

- In American public policy, ecological inferences are required to implement key features of federal law. For example, the U.S. Voting Rights Act of 1965 (and its extensions in 1970, 1975, and 1982) prohibited voting discrimination on the basis of race, color, or language. If discrimination is found, the courts or the U.S. Justice Department can order a state or local jurisdiction to redistrict its political boundaries, or to impose or prevent various other changes in electoral laws. Under present law, legally significant discrimination only exists when plaintiffs (or the Justice Department) can first demonstrate that members of a minority group (usually African American or Hispanic) vote both cohesively and differently from other voters.[6] Sometimes they must also prove that majority voters consistently prevent minorities from electing a candidate of their choice. Since survey data are rarely available in these cases, and because they are not often trustworthy in racially polarized contests, an application of the Voting Rights Act requires a valid ecological inference from electoral data and U.S. Census data.

  Voting Rights Act assessments of minority and majority voting begins with electoral returns from precincts, the smallest geographic unit for which electoral data are available. In addition to the numbers of votes received by each candidate in a precinct, census data also gives the fraction of voters in the same precinct who are African American (or other minority) or white.[7] With these two sets of aggregate data, plaintiffs must make an ecological inference about how each racial group casts its ballots. That is, since the secret ballot prevents analysts from following voters into the voting booth and peering over their shoulders as they

[6] In this book, I use "African American" and "black" interchangeably and, when appropriate or for expository simplicity, often define "white" as non-black or occasionally as a residual category such as non-black and non-Hispanic.

[7] In some states, precincts must be aggregated to a somewhat higher geographical level to match electoral and census data.

cast their ballots, the voting behavior of each racial group must be inferred using only aggregate electoral and census data. Because of the inadequacy of current methods, in some situations the wrong policies are being implemented: the wrong districts are being redrawn, and the wrong electoral laws are being changed. (Given the great importance and practicality of this problem, I will use it as a running example.)[8]

- In one election to the German Reichstag in September 1930, Adolf Hitler's previously obscure and electorally insignificant National Socialist German Worker's party became the Weimar Republic's second largest political party. The National Socialists continued their stunning electoral successes in subsequent state, local, and presidential elections, and ultimately reached 37.3% of the vote in the last election prior to their taking power. As so many have asked, how could this have happened? Who voted for the Nazis (and the other extreme groups)? Was the Nazi constituency dominated by the downwardly mobile lower middle class or was support much more widespread? Which religious groups and worker categories supported the National Socialists? Which sectors of which political parties lost votes to the Nazis? The data available to answer these questions directly include aggregate data from some of the 1,200 Kreise (districts) for which both electoral data and various census data are available. Because survey data are not available, accurate answers to these critical questions will only be possible with a valid method of ecological inference (see Hamilton, 1982; Childers, 1983; and Falter, 1991).
- Epidemiologists and public policy makers need to know whether and to what extent residential levels of radioactive radon are a risk factor for lung cancer (Stidley and Samet, 1993; Greenland and Robins, 1994a). Radon leaks through basement floors and may pose a significant health risk. Legislators in many states are considering bills that would require homeowners to test for radon and, if high levels are found, to install one of several mechanical means of reducing future exposure.

  Policymakers' decisions about such legislation obviously depend in part on the demonstrated health effects of radon. Unfortunately, collecting random samples of individual-level data would be impractical, as it would require measures of radon exposure over many years for each subject. Moreover, because only a small fraction of people with or without radon exposure get lung cancer, and because other variables like smoking are powerful covariates, reliably estimating the differences in lung cancer rates for those with different levels of radon exposure in an individual-level study would require measurements for tens of thou-

[8] The litigation based on the Voting Rights Act is vast; see Grofman, Handley, and Niemi (1992) for a review.

sands of individuals. This would be both prohibitively expensive and ethically unacceptable without altering the radon levels for individuals in a way that would probably also ruin the study. Researchers have tried case-control studies, which avoid the necessity of large samples but risk sample selection bias, and extreme-case analyses of coal miners, where the effects are larger but their high levels of radon exposure makes the results difficult to extrapolate back to residential settings. The most extensive data that remain include information such as county-level counts of lung cancer deaths from the federal Centers for Disease Control, and samples of radon concentration from each county. Ecological inferences are therefore the only hope of ascertaining the dose-response effect of radon exposure from these data. Unfortunately, without a better method of making ecological inferences, the evidence from these data will likely remain inconclusive (Lubin, 1994).[9]

- In the academic field of marketing (and its real-world counterpart), researchers try to ascertain who has bought specific products, and where advertising is most likely to be effective in influencing consumers to buy more. In many situations, researchers do not have data on the demographic and socio-economic characteristics of individuals who buy particular products, data that would effectively answer many of the research questions directly. Instead, they have extensive indirect data on the average characteristics of people in a geographic area, such as at the level of the zip code (or sometimes 9-digit zip code) in the United States. Researchers generally also have information from the company about how much of a product was sold in each of these areas. The question is, given the number of new products sold in each geographic area and, for example, the fraction of households in each area that have children, are in the upper quartile of income, are in single-parent families, or have other characteristics, how does demand for the product vary by these characteristics within each community? Only with a valid ecological inference in each geographic area can researchers learn the answers they seek. With this information, scholars will be able to study how product demand depends on these family and individual characteristics, and companies will be able to decide how to target advertising to consumers likely to be interested in their products.
- Since voter surveys are neither always possible nor necessarily reliable, candidates for political office study aggregate election returns in order

[9] Most epidemiological questions require relatively certain answers and thus, in most cases, large-scale, randomized experiments on individuals. Because each such experiment can cost hundreds of millions of dollars, a valid method of ecological inference would probably be of primary use in this field for helping scholars (and funding agencies) choose which experiments to conduct.

to decide what policies to favor, and also to tailor campaign appeals. Understanding how the support for policies varies among demographic and political groups is critical to the connections between elected officials and their constituents, and for the smooth operation of representative democracy.

- Historians are also interested in the political preferences of demographic groups, and usually for time periods for which modern survey research had not even been invented. For example, only valid ecological inferences will enable these scholars to ascertain the extent to which working-class voters supported the Socialist party in depression-era America.
- An important sociological question is the relationship between unemployment and crime, especially as affected by race and as mediated by divorce and single parenthood. Unfortunately, the best available data are usually aggregated at the level of cities or counties (Blau and Blau, 1982; Messner, 1982; Byrne and Sampson, 1986). Official U.S. government data on race-specific crime rates (in the form of the Uniform Crime Report) are usually insufficient, and individual-level survey data are in very short supply and, because they are based on self-reports, are often of dubious quality (Sampson, 1987). Only better data or a valid method of ecological inference will enable scholars to determine the critical linkages between unemployment, family disruption, race, and crime.
- The ecological inference problem, and other related aggregation problems, are central to the discipline of economics, as explained by Theil in his classic study (1954: 1): "A serious gap exists between the greater part of rigorous economic theory and the pragmatic way in which economic systems are empirically analyzed. Axiomatically founded theories refer mostly to individuals, for instance the consumer or the entrepreneur. Empirical descriptions of economic actions in large communities, on the other hand, are nearly always extremely global: they are confined to the behavior of groups of individuals. The necessity of such a procedure can scarcely be questioned.... But the introduction of relations pretending to describe the reactions of groups of individuals instead of single individuals raises questions of fundamental importance, which are not very well understood." Economists have made much progress in clarifying the links between microeconomic and macroeconomic behavior in the more than forty years since these words were written (see Stoker, 1993). They also have some good survey data, and much more impressive formal theories, but a method of ecological inference would enable economists to evaluate some of their sophisticated individual-level theoretical models more directly. This would be especially important in a field where there is much reason to value individual responses to surveys less than revealed preference measures that are best gathered at the aggregate level. Economists are also interested in developing models of

aggregate economic indicators that are built from and consistent with individual-level economic theories and data, even when the individual level is not of direct interest (see Section 14.3).

- A controversial issue in education policy is the effects of school choice voucher programs, where states or municipalities provide vouchers to students who cannot afford to attend private schools. Private schools are then composed of students from wealthy families and from those who pay with state vouchers. One of the many substantive and methodological issues in this field is determining the differential performance of students who take advantage of the voucher system to attend private schools, compared to those who would be there even without the program. Thus, data exist on aggregate school-level variables such as the dropout rate or the percent who attend college, as well as on the proportion of each private school's students who paid with a voucher. Because of privacy concerns, researchers must make ecological inferences in order to learn about the fraction of voucher students who attend college, or the fraction of non-voucher students who drop out.

The point of this list is to provide a general sense of the diversity of questions that have been addressed by (necessarily) inadequate methods of ecological inference. No tiny sample of ecological inferences such as this could do justice to the vast array of important scholarly and practical questions about individual attributes for which only aggregate data are available.

## 1.2 The Problem

On 16 and 17 November 1994, a special three-judge federal court met in Cleveland to hear arguments concerning the legality of Ohio's State House districts. A key part of the trial turned on whether African Americans vote differently from whites. Although the required facts are only knowable for individual voters, and survey data were unavailable (and are unreliable in the context of racial politics), the only relevant information available to study this question was political and demographic data at the aggregate level.[10]

Table 1.1 portrays the issue in this case as an example of the more general ecological inference problem. This table depicts what is known

[10] I had a small role in this case as a consultant to the state of Ohio and therefore witnessed the following story firsthand. My primary task in the case was to evaluate the relative fairness of the state's redistricting plan to the political parties, using methods developed in King and Browning (1987), King (1989b), and Gelman and King (1990, 1994a, b).

| Race of Voting-Age Person | Voting Decision | | | |
|---|---|---|---|---|
| | Democrat | Republican | No Vote | |
| black | ? | ? | ? | 55,054 |
| white | ? | ? | ? | 25,706 |
| | 19,896 | 10,936 | 49,928 | 80,760 |

Table 1.1 The Ecological Inference Problem at the District Level: The 1990 Election to the Ohio State House, District 42. The goal is to infer from the marginal entries (each of which is the sum of the corresponding row or column) to the cell entries.

for the election to the Ohio State House that occurred in District 42 in 1990. The black Democratic candidate received 19,896 votes (65% of votes cast) in a race against a white Republican opponent. African Americans constituted 55,054 of the 80,760 people of voting age in this district (68%). Because this known information appears in the margins of the cross-tabulation, it is usually referred to as the *marginals*. The ecological inference problem involves replacing the question marks in the body of this table with inferences based on information from the marginals. (Ecological inference is traditionally defined in terms of a table like this and thus in terms of discrete individual-level variables. Most political scientists, sociologists, and geographers, and some statisticians, have retained this original definition. Epidemiologists and some others generalize the term to include any aggregation problem, including continuous individual-level variables. I use the traditional definition in this book in order to emphasize the distinctive characteristics of aggregated discrete data, and discuss aggregation problems involving continuous individual-level variables in Chapter 14.)

For example, the question mark in the upper left corner of the table represents the (unknown) number of blacks who voted for the Democratic candidate. Obviously, a wide range of different numbers could be put in this cell of the table without contradicting its row and column marginals, in this case any number between 0 and 19,896, a logic referred to in the literature as *the method of bounds*.[11] As a result, some other information or method must be used to further narrow the range of results.

[11] That is, although the row total is 55,054, the total number of people in the upper left cell of Table 1.1 cannot exceed 19,896, or it would contradict its column marginal.

| Race of Voting-Age Person | Voting Decision | | | |
|---|---|---|---|---|
| | Democrat | Republican | No Vote | |
| black | ? | ? | ? | 221 |
| white | ? | ? | ? | 484 |
| | 130 | 92 | 483 | 705 |

Table 1.2 The Ecological Inference Problem at the Precinct Level: Precinct P in District 42 (1 of 131 in the district described in Table 1.1). The goal is to infer from the margins of a set of tables like this one to the cell entries in each.

Fortunately, somewhat more information is available in this example, since the parties in the Ohio case had data at the level of precincts (or sometimes slightly higher levels of aggregation instead, which I also will refer to as precincts). Ohio State House District 42 is composed of 131 precincts, for which information analogous to Table 1.1 is available. For example, Table 1.2 displays the information from Precinct P, which in District 42 falls between Cascade Valley Park and North High School in the First Ward in the city of Akron. The sum of any item in the precinct tables, across all precincts, would equal the number in the same position in the district table. For example, if the number of blacks voting for the Democratic candidate in Precinct P were added to the same number from each of the other 130 precincts, we would arrive at the total number of blacks casting ballots for the Democratic candidate represented as the first cell in Table 1.1.

The ecological inference problem does not vanish by having access to the precinct-level data, such as that in Table 1.2, because we ultimately require individual-level information. Each of the cells in this table is still unknown. Thus, knowing the parts would tell us about the whole, but disaggregation to precincts does not appear to reveal much more about the parts.

With a few minor exceptions, no method has even been proposed to fill in the unknown quantities at the precinct level in Table 1.2. What scholars have done is to develop methods to use the observed variation in the marginals over precincts to help narrow the range of results at the district level in Table 1.1. For example, if the Democratic candidate receives the most votes in precincts with the largest fractions of African Americans, then it seems intuitively reasonable to suppose that blacks are voting disproportionately for the Democrats (and thus

the upper left cell in Table 1.1 is probably large). This assumption is often reasonable, but Robinson showed that it can be dead wrong: the individual-level relationship is often the opposite sign of this aggregate correlation, as will occur if, for example, whites in heavily black areas tend to vote more Democratic than whites living in predominately white neighborhoods.

Unfortunately, even the best available current methods of ecological inference are often wildly inaccurate. For example, at the federal trial in Ohio (and in formal sworn deposition and in a prepared report), the expert witness testifying for the plaintiffs reported that 109.63% of blacks voted for the Democratic candidate in District 42 in 1990! He also reported in a separate, but obviously related, statement that a negative number of blacks voted for the Republican candidate. Lest this seem like one wayward result chosen selectively from a sea of valid inferences, consider a list of the results from all districts reported by this witness (every white Republican who faced a black Democrat since 1986), which I present in Table 1.3. A majority of these results are over 100%, and thus impossible. No one was accusing the Democratic candidates of stuffing the ballot box; dead voters were not suspected of turning out to vote more than they usually do. Rather, these results point out the failure of the general methodological approach. For those familiar with existing ecological inference methods, these results may be disheartening, but they will not be surprising: impossible results occur with regularity.

What of the analyses in Table 1.3 that produced results that were not impossible? For example, in District 25, the application of this standard method of ecological inference indicated that 99% of blacks voted for the Democratic candidate in 1990. Is this correct? Since no external information is available, we have no idea. However, we do know, from other situations where data do exist with which to verify the results of ecological analyses, that the methods usually do not work. The problem, of course, is that when they give results that are technically possible we might be lulled into believing them. As Robinson so clearly stated, even technically possible results from these standard methods are usually wrong.

When ridiculous results appear in academic work, as they sometimes do, there are few practical ramifications. In contrast, inaccurate results used in making public policy can have far-reaching consequences. Thus, in order to attempt to avoid this situation, the witness in this case used the best available methods at the time and had at his disposal far more resources and time than one would have for almost any academic project. The partisan control of a state legislature was at stake, and research resources were the last things that would be

| *Year* | *District* | *Estimated Percent of Blacks Voting for the Democratic Candidate* |
|---|---|---|
| 1986 | 12 | 95.65% |
| | 23 | 100.06 |
| | 29 | 103.47 |
| | 31 | 98.92 |
| | 42 | 108.41 |
| | 45 | 93.58 |
| 1988 | 12 | 95.67 |
| | 23 | 102.64 |
| | 29 | 105.00 |
| | 31 | 100.20 |
| | 42 | 111.05 |
| | 45 | 97.49 |
| 1990 | 12 | 94.79 |
| | 14 | 97.83 |
| | 16 | 94.36 |
| | 23 | 101.09 |
| | 25 | 98.83 |
| | 29 | 103.42 |
| | 31 | 102.17 |
| | 36 | 101.35 |
| | 37 | 101.39 |
| | 42 | 109.63 |
| | 45 | 97.62 |

Table 1.3 Sample Ecological Inferences: All Ohio State House Districts Where an African American Democrat Ran Against a White Republican, 1986–1990. *Source*: "Statement of Gordon G. Henderson," presented as part of an exhibit in federal court. Figures above 100% are logically impossible.

spared if the case could be won. (The witness also had extensive experience testifying in similar cases.) Moreover, he was using a method (a version of Goodman's "ecological regression") that the U.S. Supreme Court had previously declared to be appropriate in applications such as this (*Thornburg v. Gingles*, 1986). If there was any way of avoiding these silly conclusions, he certainly would have done so. Yet, even with all this going for him he was effectively forced by the lack of better methods to present results that indicated, in over half the districts he studied, that more African Americans voted for the Democratic candidate than there were African Americans who voted.

Two types of statistical difficulties cause inaccurate results such as these in ecological inferences. The first is *aggregation bias*. This is the

effect of the information loss that occurs when individual-level data are aggregated into the observed marginals. The problem is that in some aggregate data collections, the type of information loss may be selective, so that inferences that do not take this into account will be biased.

The second cause of inaccurate results in ecological inferences is a variety of *basic statistical problems*, unrelated to aggregation bias, that have not been incorporated into existing methods. These are the kinds of issues that would be resolved first in any other methodological area, although most have not yet been addressed. For example, much data used for ecological inferences have massive levels of "heteroskedasticity" (a basic problem in regression analysis), but this has never been noted in the literature—and sometimes explicitly denied—even though it is obviously present even in most published scatter plots (about which more in Chapter 4).

## 1.3 The Solution

This section sets forth seven characteristics of the proposed solution to the ecological inference problem not met by previous methods. However, unlike the proof of a mathematical theorem, statistical solutions can usually be improved continually—hence the phrase *a* solution, rather than *the* solution, in the title of this book. Modern statistical theory does not date back even as far as the ecological inference problem, so as we learn more we should be able to improve on this solution further. Similarly, as computers continue to get faster, we can posit more sophisticated models that incorporate more information. The method offered here is the first that consistently works in practice, but it is also intended to put the ecological inference literature on a firmer theoretical and empirical foundation, helping to lead to further improvements.

First, *the solution is scientifically validated with real data*. Several extensive collections of real aggregate data, for which the inner cells of the cross-tabulation are known from public records, are used to help validate the method. For example, estimates of the levels of black and white voter registration are compared to the known answer in public records. (These are real issues, not contrived for the purpose of a methodological treatise; they are the subject of considerable academic inquiry, and even much litigation in many states.) Data from the U.S. Census aggregated to precinct-sized aggregates in South Carolina are used to study the relative frequency with which males and females are in poverty. Also useful for this purpose are data from Atlanta, Georgia, that include information about voter loyalty and defection rates

in the transitions between elections, and from turn-of-the-century U.S. county-level data on black and white literacy rates, in order to validate the model in those contexts. Finally, I have been able to study the properties of aggregate data extensively with a large collection of merged U.S. Census data and precinct-level aggregate election data for most electoral offices and the entire nation. The method works in practice. In contrast, if the only goal were to develop a method that worked merely in theory, then the problem might already have been considered "solved" long ago, as the literature includes many methods that work only if a list of unverifiable assumptions are met.

Using data to evaluate methodological approaches is, of course, good scientific practice, but it has been rare in this field that has focused so exclusively on hypothetical data, and on theoretical arguments without economic, political, sociological, psychological, or other foundations. Indeed, the entire ecological inference literature contains only forty-nine comparisons between estimates from aggregate data and the known true individual-level answer.[12] (Because this work includes a variety of new data sets, and a method that gives

[12] This estimate of the number of times authors in the ecological inference literature have made themselves vulnerable to being wrong is based on counting data sets original to this literature. Individual cross-tabulations that were used to study the method of bounds are excluded since no uncertainty, and thus no vulnerability, exists. I obviously also exclude studies that use data sets previously introduced to this literature. A list of data sets and the studies in which they were first used are as follows: Race and illiteracy from the 1930 U.S. Census (Robinson, 1950); race by domestic service from community area data (Goodman, 1959; used originally to study bounds by Duncan and Davis, 1953); infant mortality by race and by urbanicity in U.S. states (Duncan et al., 1961: 71–72); 1964–1966 voter transitions in British constituencies (Hawkes, 1969); a voter transition between Democratic primaries in Florida (Irwin and Meeter, 1969); a 1961 German survey (Stokes, 1969); voter transition in England from Butler and Stokes (1969) data (Miller, 1972); survey of first-year university students (Hannan and Burstein, 1974); vote for Labour by worker category (Crewe and Payne, 1976); voter transition in England compared to a poll (McCarthy and Ryan, 1977); voter transition February to October 1974 in England compared to a poll (Upton, 1978); voter transition from a general election in 1983 to an election to the European parliament in 1984 compared to an ITN poll (Brown and Payne, 1986); one comparison based on twenty-four observations from Lee County, South Carolina, comparing registration and turnout by race (Loewen and Grofman, 1989); two comparisons of a survey to Swedish election data (Ersson and Wörlund, 1990); twenty comparisons of aggregate electoral data in California and nationally compared to exit polls, comparisons using census data, and official data on registration and voter turnout (Freedman et al., 1991); eight voter transition studies in Denmark compared to survey data (Thomsen et al., 1991); race and registration data from Matthews and Prothro (1966) (Alt, 1993); race and literacy from the 1910 U.S. Census (Palmquist, 1994); housing tenure transitions from 1971 to 1981 in England from census data (Cleave, Brown, and Payne, 1995). If you know of any work that belongs on this list that I missed, I would appreciate hearing from you.

district- and precinct-level estimates, the book presents over sixteen thousand such comparisons between estimates and the truth.) Many of these forty-nine ecological inferences are compared to estimates from sample surveys, but scholars rarely correct for known survey biases with post-stratification or other methods.[13] Others use "data" that are made up by the investigator, such as those created with computerized random number generators. All these data sets have their place (and some will have their place here too), but their artificial nature, exclusive use, and especially limited number and diversity fail to present the methodologist with the kinds of problems that arise in using real aggregate data and studying authentic social science problems. Scholars are therefore unable to adapt the methods to the opportunities in the data and will not know how to avoid the likely pitfalls that commonly arise in practice.

Second, *the method described here offers realistic assessments of the uncertainty of ecological estimates*. Reporting the uncertainty of one's conclusions is one of the hallmarks of modern statistics, but it is an especially important problem here. The reason is that ecological inference is an unusual statistical problem in which, under normal circumstances, we never observe realizations of our quantity of interest. For example, since most German citizens who voted for the Nazi party are no longer around to answer hypothetical survey questions, and could hardly be expected to answer them sincerely even if they were, no method will ever be able to fill in the cross-tabulation with certainty. Thus a key component of any solution to this problem is that correct uncertainty estimates be an integral part of all inferences.

Many methods proposed in the literature provide no uncertainty estimates. Others give uncertainty estimates that are usually incorrect (as for example when 95% confidence intervals do not capture the correct answer about 95% of the time). The method proposed here provides reasonably accurate (and empirically verified) uncertainty estimates. Moreover, these estimates are useful since the intervals turn out to be narrower than one might think.

Third, *the basic model is robust to aggregation bias*. Although this book also includes modifications of this basic model to compensate for aggregation bias explicitly, these modifications are often unnecessary. That is, even when the process of aggregation causes existing methods to give answers that bear no relationship to the truth, the method proposed here still usually gives accurate answers.

[13] Surveys are also very underused in this literature, perhaps in part since many scholars came to this field because of their skepticism of public opinion polls.

In order to develop an explicit approach to avoiding aggregation bias, I prove that the numerous and apparently conflicting explanations for aggregation bias are mathematically equivalent, even though they each appear to offer very different substantive insights. This theoretical result eliminates the basis for existing scholarly disagreements over which approach is better, or how many problems we need to deal with. All problems identified with aggregation bias are identical; only one problem needs to be solved. In the cases where an explicit treatment of aggregation bias is necessary under the proposed model, this result makes possible the model generalization required to accomplish the task.

Fourth, *all components of the proposed model are in large part verifiable in aggregate data*. That is, although information is lost in the process of aggregation, and thus ecological inferences will always involve risk, some observable implications of all model assumptions remain in aggregate data. These implications are used to develop diagnostic tests to evaluate the appropriateness of the model to each application, and to develop generalizations for the times when the assumptions of the basic model are contradicted by the data. Thus, the assumptions on which this model is based can usually be verified in sufficient detail in aggregate data in order to avoid problems that cause other methods to lose their bearing.

Fifth, *the solution offered here corrects for a variety of serious statistical problems, unrelated to aggregation bias, that also affect ecological inferences*. It explicitly models the main source of heteroskedasticity in aggregate data, allows precinct-level parameters to vary, and otherwise includes far more known information in the model about the problem.

The sometimes fierce debates between proponents of the deterministic "method of bounds" and supporters of various statistical approaches are resolved by combining their (largely noncontradictory) insights into a single model. Including the precinct-level bounds in the statistical model substantially increases the amount of information used in making ecological inferences. For example, imagine that every time you run a regression, you could take some feature of the model (such as a predicted value), hold it outside a window and, if it is wrong—completely wrong with no uncertainty—the clouds would part and a thunderbolt would turn your computer printout into a fiery crisp. Remarkably, although they have not been exploited in previous statistical models, the bounds provide exactly this kind of certain information in all ecological inference problems for each and every observation in a data set (albeit perhaps with a bit less fanfare). In any other field of statistical analysis, this valuable information, and the other more ordinary statistical problems, would be

addressed first, and yet most have been ignored. Correcting these basic statistical problems is also what makes this model robust to aggregation bias.

Sixth, *the method provides accurate estimates not only of the cells of the cross-tabulation at the level of the district-wide or state-wide aggregates but also at the precinct level*. For example, the method enables one to fill in not only Table 1.1 with figures such as the fraction of blacks voting for the Democrats in the entire district, but also the precinct-level fractions for each of the 131 tables corresponding to Table 1.2. This has the obvious advantage of providing far more information to the analyst, information that can be studied, plotted on geographic maps, or used as dependent variables in subsequent analyses. It is also quite advantageous for verifying the method, since 131 tests of the model for each data set are considerably more informative than one.

Finally, *the solution to the ecological inference problem turns out to be a solution to what geographers' call the "modifiable areal unit problem."* The modifiable areal unit problem occurs if widely varying estimates result when most methods are applied to alternate reaggregations of the same geographic (or "areal") units. This is a major concern in geography and related fields, where numerous articles have been written that rearrange geographic boundaries only to find that correlation coefficients and other statistics totally change substantive interpretations (see Openshaw, 1979, 1984; Fotheringham and Wong, 1991). In contrast, the method given here is almost invariant to the configuration of district lines. If precinct boundaries were redrawn, even in some random fashion, inferences about the cells of Table 1.1 would not drastically change in most cases.

Every methodologist dreams of inventing a statistical procedure that will work even if the researcher applying it does not understand the procedure or possess much "local knowledge" about the substance of the problem. This dream has never been fulfilled in statistics, and the same qualification holds for the method proposed here: The more contextual knowledge a researcher makes use of, the more likely the ecological inference is to be valid. The method gives the researcher with this local knowledge the tools to make a valid ecological inference. That is, with a fixed, even inadequate, amount of local knowledge about a problem, a researcher will almost always do far better by using this method than those previously proposed. But making valid ecological inferences is not usually possible without operator intervention. Valid inferences require that the diagnostic tests described be used to verify that the model fits the data and that the distributional assumptions apply. Because the basic problem is a lack of information,

bringing diverse sources of knowledge to bear on ecological inferences can have an especially large payoff.

## 1.4 The Evidence

As a preview of Part IV, which reports extensive evaluations of the model from a variety of data sets, this section gives just two applications, one to demonstrate the accuracy of the method and the other to portray how much more information it reveals about the problem under study. The first application provides 3,262 evaluations of the ecological inference model presented in this book—67 times as many comparisons between estimates from an aggregate model and the truth as exist in the entire history of ecological inference research. The second is a brief geographic analysis in another application that serves to emphasize how much more information about individual behavior this method provides than even the (unrealized) goal of previous methods.

The data for the first application come from the state of Louisiana, which records by precinct the number of blacks who vote and the number of whites who vote (among those registered). These data make it possible to evaluate the ecological inference model described in this book as follows. For each of Louisiana's 3,262 precincts, the procedure uses only aggregate data: the fraction of those registered who are black and the fraction of registered people turning out to vote for the 1990 elections (as well as the number registered). These aggregate, precinct-level data are then used to estimate the fraction of blacks who vote in each precinct. Finally, I validate the model by comparing these estimates to the true fractions of blacks who turn out to vote. (That is, the true fractions of black and white turnout are not used in the estimation procedure.)[14]

One brief summary of the results of this analysis appears in Figure 1.1. This figure plots the estimated fraction of blacks turning out to vote in 1990 (horizontally) by the true fraction of blacks voting in that year (vertically). Each precinct is represented in the figure by a circle with area proportional to the number of blacks in the precinct. If the model estimates were exactly correct in every precinct, each

[14] The 3,262 evaluations of the model in this section are from the same data set and, as such, are obviously related. However, each comparison between the truth and an estimate provides a separate instance in which the model is vulnerable to being wrong. These model evaluations simulate the usual situation in which the ecological analyst has no definite prior knowledge about whether the parameters of interest are dependent, unrelated, or all identical.

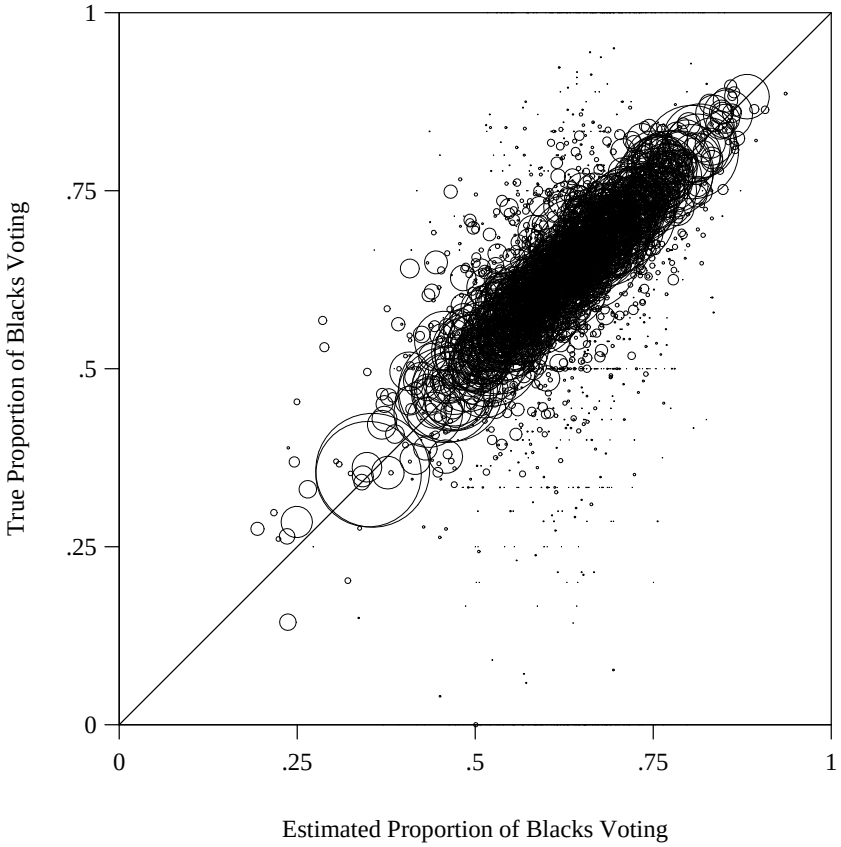


Figure 1.1 Model Verification: Voter Turnout among African Americans in Louisiana Precincts. This figure represents 3,262 precincts in 1990, with each circle size proportional to the number of voting-age African Americans in the precinct. That the vast majority of circles fall near the diagonal line, indicating that the estimated and true fractions of blacks voting are nearly identical, is strong confirmation of the model.

circle would be centered exactly on the 45° line. In fact, almost all of the 3,262 precincts fall on or near this diagonal line, demonstrating the success of this method of making inferences about individual behavior using only aggregate data. The few precincts that are farther from the line have tiny numbers of African Americans, so the vast majority of individual voters are correctly estimated.

The results are compelling. If Figure 1.1 were merely a plot of the observed values of a variable by the fitted values of the same vari-

able used during the estimation procedure, any empirical researcher should be pleased: the fit is extremely good. If instead the figure were based on the harder problem of making out-of-sample predictions, where past realizations were used to calibrate the prediction, the result would be even better. But the result here is even more dramatic, since the estimates in the figure were computed from only aggregate data. The true fraction of blacks turning out to vote (the vertical dimension in the figure) was not part of the estimation procedure. Moreover, no past realizations of the truth being estimated were used.

Part IV provides many more model evaluations and of many types. These evaluations include data sets for which existing methods do reasonably well at estimating the statewide average, in which case the method offered here also gives reasonable statewide results and in addition much more information in the form of correct confidence intervals and accurate results for each precinct in the state. Part IV also gives examples of data sets where existing methods are hopelessly biased, but the method offered here gives highly accurate estimates. For example, the best existing method indicates that 20% fewer males in South Carolina fall below the poverty level than there are males in that state (see Table 11.2 on page 220). In contrast, the method offered here gives accurate answers for this statewide aggregate (see Figure 11.2 on page 222) as well as for the fraction of males in poverty in each of the 3,187 precinct-sized geographic units (see Figure 11.3 on page 223).

The book also includes situations in which almost all information was aggregated away and standard methods give even more ridiculous results; in those cases, the method described here gives reasonable results with wider confidence intervals, reflecting accurately the degree of uncertainty in the ecological inference (see Chapter 12). The method usually even gives accurate estimates when all the conditions for "aggregation bias" are met, when the process of aggregation eliminates most of the variation in one of the aggregate variables, and when extrapolations far from the range of observed data are necessary. In all these difficult examples, the method offered here gives accurate answers with correct confidence intervals. The method will not always work: since information is lost during aggregation, no method of ecological inference could work in all data sets. However, the procedures introduced here come with diagnostics that researchers can use to evaluate the risks and avoid the problems in most cases.

Finally, I give a brief report of an analysis of 1990 turnout by race in New Jersey's 567 minor civil divisions (mostly cities and towns). These data cannot be used to verify ecological inferences since the true individual-level answers are not known, but they can be used

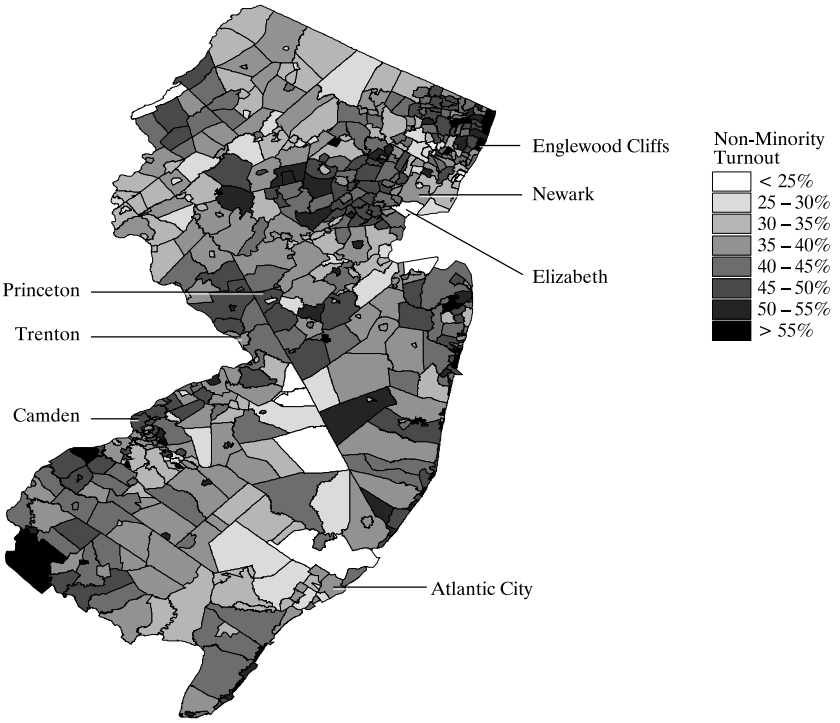


Figure 1.2 Non-Minority Turnout in New Jersey Cities and Towns. In contrast to the best existing methods, which provide *one* (incorrect) number for the entire state, the method offered here gives an accurate estimate of white turnout for all 567 minor civil divisions in the state, a few of which are labeled.

to demonstrate how much more information the method offered here provides to users. The most popular existing method (Goodman's regression) gives only two numbers of relevance, the state-wide fractions of blacks who vote and whites who vote (the latter estimate, incidentally, is five standard deviations above its maximum possible value given by the method of bounds). In contrast, the solution to the ecological inference problem offered here gives reliable estimates of these two numbers for the state-wide average as well as for each of the 567 cities and towns.

In order to emphasize the rich information this method unearths, Figure 1.2 maps the estimated degree of voter turnout among non-minorities. In this map, minor civil divisons in New Jersey are given darker shades when the estimated degree of non-minority voter turnout is higher. A few landmarks are labeled to give readers some

bearing. The vast increase in information the method provides is represented by the interesting geographic variation in this map (and an *additional* complete map for minority turnout). For example, Figure 1.2 shows that non-minority turnout is substantially higher in the city of Newark than the neighboring city of Elizabeth. Is this because of a racial threat posed by Newark's larger minority population? Is the white mobilization in the wealthy towns of Bergen County near Englewood Cliffs a result of the state government's attempt to integrate schools by regionalizing its school districts? By providing reliable individual-level geographic-based information, the solution to the ecological inference problem can be used to raise numerous questions such as these. The method also provides opportunities for answering such questions by using the estimates provided as dependent variables in second-stage analyses (using, in this case, explanatory variables such as fraction minority population, or state attempts at integration).

## 1.5 The Method

This section gives a brief non-mathematical sketch of the nature of the basic model introduced. Although several approaches are discussed in the methodological literature, the only method of ecological inference widely used in practice is Goodman's model, which is based on a straightforward linear regression and effectively assumes that the quantities of interest (such as the proportion of blacks and whites who vote) are constant over all precincts (see Section 3.1). Allowing these quantities to vary over the precincts and estimating them all, as is done in this book, provides far more detailed information about the individual-level relationships, and moderately improves the overall results.

Applying the deterministic information from the method of bounds to each and every precinct-level quantity of interest provides very substantial improvements and makes inferences especially robust to aggregation bias. Goodman's regression does not restrict the quantities of interest (which are proportions) even to the [0,1] interval. Many have suggested modifying Goodman's regression by restricting these aggregate quantities of interest to this interval, but this results in implausible corner solutions and, more importantly, imposes no restrictions on any of the individual precinct quantities. In contrast, the method offered here uses the bounds on the quantities of interest in every precinct, most of which turn out in practice to be much narrower than [0,1]. Because, also, these bounds are known with cer-

tainty, this procedure adds a surprising amount of information to the statistical model.[15]

This combination of the precinct-level deterministic bounds with a statistical model unifies the two primary competing parts of the ecological inference literature. First, by treating each precinct in isolation, the method uses all available information to give a range of *possible* values for its precinct-level quantities of interest. Then, in order to close in further on the right answer, the statistical model "borrows strength" from all the other precincts in the data set to give the *probable* location of each true quantity of interest within its known deterministic bounds.

The method introduced also includes a model of variability that matches the patterns in real aggregate data and that is internally consistent even in the presence of areal units that are modified. This and other features provide another significant boost in the performance of the model. Extensions of the model allow for the model assumptions to be evaluated, modified, or dropped, and for several types of external information to be included. A fully nonparametric version is also provided.

Some features of the model are related in part to variable parameter models in econometrics (e.g., Swamy, 1971); empirical Bayesian models in statistics and biostatistics (Efron and Morris, 1973; Rubin, 1980; Breslow, 1990); Manski's (1995) approach to identification via parameter bounds; models of multiple imputation for missing values in surveys (Rubin, 1987) and for coarse data problems (Heitjan, 1989; Heitjan and Rubin, 1990); hierarchical linear models in education research (Bryk and Raudenbush, 1992); and "inverse problems" in tomographic imaging (Vardi et al., 1985; Johnstone and Silverman, 1990). The solution to the ecological inference problem offered here is also related to some statistical models for the aggregation of individual-level continuous variables developed in econometrics (Stoker, 1993), as described in Section 14.3.

[15] As an analogy, consider how much information could be added to the usual linear regression if we knew for certain a different narrow range within which each observation's $\hat{y}$ must fall.

CHAPTER 2

# Formal Statement of the Problem

THIS CHAPTER formalizes the ecological inference problem as introduced in Chapter 1. It provides notation that will be used throughout the rest of the book and identifies the quantities of interest at each level of analysis (see also the Glossary of Symbols at page 313).

The ecological inference problem begins with a set of cross-tabulations for each of $p$ aggregate units. Given the marginals from each of the $p$ tables, the goal is to make inferences about the cells of each of the tables. The $p$ cross-tabulations are usually from geographic units, such as precincts districts, or counties.[1] For electoral applications, choosing data in which all geographic units have the same candidates (such as precincts from the same district or counties from the same statewide election) is advisable so that election effects are controlled. The cross-tabulations could also be groups of survey respondents (such as the fractions of working-class and middle-class voters preferring the Labour party) in a series of independent cross-sections and for which we wish to estimate the transitions between groups.

All results and models in this book can be generalized to arbitrarily large contingency tables, as demonstrated in Section 8.4 and Chapter 15. The method of ecological inference introduced is also applicable to almost all types of aggregate data, and is not limited by substantive area. However, the method is capable of taking advantage of whatever additional substantive information is available about a specific ecological inference. In order to highlight the types of information to watch out for, I introduce the notation in this chapter and the model in the rest of the book in the context of a specific substantive example. This will also fix ideas and make it easier to follow the subsequent algebraic developments. The example causes no loss of generality, even though all applications have unique elements. The specific example is based on various aspects of race and voting, as introduced in Chapter 1. This example is real and has important practical, scholarly, legal, and public policy implications. Details about it appear throughout the book in order to give a better sense of how the arguments

[1] Wherever possible, I use notation that is mnemonic, and identify this in the text by underlying the relevant character in the corresponding word. Thus, in this case, $p$ is mnemonic for precincts.

| Race of Voting-Age Person | Voting Decision: Democrat | Republican | No Vote | |
|---|---|---|---|---|
| black | $N_i^{bD}$ | $N_i^{bR}$ | $N_i^{bN}$ | $N_i^b$ |
| white | $N_i^{wD}$ | $N_i^{wR}$ | $N_i^{wN}$ | $N_i^w$ |
| | $N_i^D$ | $N_i^R$ | $N_i^N$ | $N_i$ |

Table 2.1 Basic Notation for Precinct $i$, $i = 1, \ldots, p$. All items in this table are *counts* of the Number of people in each cell position. The two elements in each superscript refer to the row and column position, respectively, with mnemonics indicated by the underlined letter in the labels. The column and row marginals, which are sums of the elements in the corresponding row or column, are observed. The interior cell entries are the object of inference.

apply to similar problems in other empirical examples. Part IV analyzes real data from this particular example, and from a diverse variety of others.

Begin by delineating a formal version of Table 1.2 (page 14) for each of $p$ individual precincts.[2] Table 2.1 provides some notation for observed data and unobserved quantities of interest.

This table describes a single precinct (or other geographic entity) labeled $i$ from a data set of $p$ precincts in a single electoral district (such as a state assembly seat). The table is based on a simple example with two variables, the race of the voting-age person (black or white, with "white" defined as non-black) and the voting decision (Democrat, Republican, or no vote).

Every symbol in Table 2.1 has a subscript $i$, referring to the $i^{\text{th}}$ precinct ($i = 1, \ldots, p$). Each cell in Table 2.1 is a raw count of the number of people who fall in that cell. Superscripts refer to positions in the table (and thus values of the row and column variables, respectively). For example, $N_i^{bD}$ is the Number of black persons of voting age casting a ballot for the Democratic candidate in precinct $i$. I denote

[2] In order to gather both race and electoral results, electoral precincts must be matched with census geography. This sometimes means that precincts must be aggregated to a slightly higher level. The Census Bureau calls these "voter tabulation districts" or VTDs, although it sometimes makes sense to use "places," "minor civil divisions," counties, school districts, or other census jurisdictions. I use the more familiar term "precincts" to refer to the lowest level of geography for which both variables can be collected within an electoral district.

| Race of Voting-Age Person | Voting Decision: Democrat | Republican | *Subtotal* (*Turnout*) | No Vote | |
|---|---|---|---|---|---|
| black | $\lambda_i^b$ | $1-\lambda_i^b$ | $\beta_i^b$ | $1-\beta_i^b$ | $X_i$ |
| white | $\lambda_i^w$ | $1-\lambda_i^w$ | $\beta_i^w$ | $1-\beta_i^w$ | $1-X_i$ |
| | $D_i$ | | $T_i$ | $1-T_i$ | |

Table 2.2 Alternative Notation for Precinct $i$. This table reexpresses the elements of Table 2.1 as *proportions*, and inserts an extra summary column for voter turnout. The goal is to estimate the quantities of interest, the fraction of blacks and whites who vote ($\beta_i^b$, $\beta_i^w$) and who vote for the Democratic candidate ($\lambda_i^b$, $\lambda_i^w$), from the aggregate variables, the fraction of voting-age people who are black ($X_i$), who vote ($T_i$), and who vote for the Democrat ($D_i$), along with the number of voting-age people ($N_i$).

aggregation by dropping the superscript or subscript corresponding to the dimension being summed. This includes column totals (such as the number of Republicans in precinct $i$, $N_i^R = N_i^{bR} + N_i^{wR}$), row totals (such as the number of blacks in precinct $i$, $N_i^b = N_i^{bD} + N_i^{bR} + N_i^{bN}$), and the number of voting-age people in the entire precinct (as indicated by the symbol in the bottom right corner of the table, $N_i$).

Although the basic ecological inference problem is described completely in Table 2.1, the following summaries of it will prove convenient for later analysis. First, denote the total number of blacks who Turn out to vote as $N_i^{bT} = N_i^{bD} + N_i^{bR}$, whites who Turn out as $N_i^{wT}$, and total Turnout as $N_i^T$. Then, Table 2.2 reexpresses all the counts as proportions, and also inserts a subtotal column between the "Republican" and "No vote" columns to refer to voter turnout proportions. The meaning of the proportion in the enclosed box in Table 2.2 corresponds to the count in the same position of each enclosed box in Table 2.1.

Table 2.3 is the final table of notation, and it is taken from the last three columns of Table 2.2. Whenever possible, this simpler $2 \times 2$ table serves as our running example, with variables black vs. white, and vote vs. no vote.

The key to the ecological inference problem is that researchers only observe the marginals in these tables—the final row (summarized by $D_i$ and $T_i$) and final column (summarized by $X_i$), along with $N_i$:

$D_i$ Proportion of voting-age population choosing the Democratic candidate, $N_i^D/N_i$

| Race of Voting-Age Person | Voting Decision: Vote | Voting Decision: No Vote | |
|---|---|---|---|
| black | $\beta_i^b$ | $1-\beta_i^b$ | $X_i$ |
| white | $\beta_i^w$ | $1-\beta_i^w$ | $1-X_i$ |
| | $T_i$ | $1-T_i$ | |

Table 2.3 Simplified Notation for Precinct $i$. This table was formed from the rightmost three columns of Table 2.2. The goal is to infer the quantities of interest, $\beta_i^b$ (the fraction of blacks who vote) and $\beta_i^w$ (the fraction of whites who vote), from the aggregate variables $X_i$ (the fraction of voting-age people who are black) and $T_i$ (the fraction of people who vote), along with $N_i$ (the number of voting-age people).

$T_i$ Proportion of voting-age population Turning out to vote, $N_i^T/N_i$
$X_i$ Proportion of voting-age population who are black, $N_i^b/N_i$, an eXplanatory variable
$N_i$ Number of people of voting age

The goal of ecological inference is to learn about the quantities within the body of the table on the basis of the information from the margins—to learn about the Greek letters, representing information at the individual level, from the information in the Roman letters, which stand for the aggregate data. The quantities of interest can be summarized by four parameters defined for each precinct $i$:

$\beta_i^b$ Proportion of voting-age blacks who vote, $N_i^{bT}/N_i^b$
$\beta_i^w$ Proportion of voting-age whites who vote, $N_i^{wT}/N_i^w$
$\lambda_i^b$ Proportion of black voters choosing the Democratic candidate, $N_i^{bD}/N_i^{bT}$
$\lambda_i^w$ Proportion of white voters choosing the Democratic candidate, $N_i^{wD}/N_i^{wT}$

When focusing on the pared-down Table 2.3, $\beta_i^b$ and $\beta_i^w$ are the only parameters of interest, and $X_i$, $T_i$, and $N_i$ are the observed aggregate marginals.

Although the ultimate goal of ecological inference, and the problem solved in this book, is learning about these precinct-level parameters, virtually all previous scholars have limited their inquiry to learning about the quantities of interest averaged over all people in the voting-age population in the entire district. These aggregates may be obtained

from the table either directly, by applying the precinct formulas to the district totals, or indirectly, by taking a *weighted* average of all $p$ precinct parameters, where the weights are functions of the precincts' black or white voting-age populations. For example, the fraction of blacks voting in the entire district is computed either directly,

$$B^b = \frac{\sum_{i=1}^{p} N_i^{bT}}{N^b}$$

or indirectly,

$$= \frac{\sum_{i=1}^{p} N_i^b \beta_i^b}{N^b}$$

where the number of blacks of voting age in the entire district (that is, in all $p$ precincts) is $N^b = \sum_{i=1}^{p} N_i^b$. The equivalence of these two expressions is obvious as expressed here (since $\beta_i^b = N_i^{bT}/N_i^b$), even though in the literature this weighted average is often confused with the unweighted average, which I denote as $\mathfrak{B}^b = \frac{1}{p}\sum_{i=1}^{p} \beta_i^b$. Since most analyses will be in terms of the precinct parameters, the appropriate weights are very important if interest shifts to the district-level parameters (about which more in Chapter 4).

The problem with ignoring the difference between the weighted $B^b$ and unweighted $\mathfrak{B}^b$ is what we might call the *Manhattan Effect* due to this simple example: Suppose a researcher wishes to make an ecological inference about the fraction of blacks who support each candidate in a mayoral election in New York City. Because of the difficulties of matching electoral precincts and census geography in Manhattan (the largest of New York's five boroughs), it can not be broken down into smaller aggregate units, even though the rest of the city is broken into numerous precinct-sized units of about 700 people each. The problem is not only that that Manhattan's population is massive compared to any of the other units, but that it frequently votes differently from the rest of the city. Thus, weighting Manhattan's votes in making ecological inferences as equivalent to one 700-person precinct would discard an enormous amount of information and wreak havoc on any estimates of the city-wide proportion of blacks who vote for each candidate. The solution to the Manhattan Effect is to take into account the size of the population of each aggregate unit and to compute the weighted ($B^b$) rather than unweighted ($\mathfrak{B}^b$) average of the $\beta_i^b$'s.

The four aggregate parameters of interest (using the corresponding upper case Greek letter in each case) include the district-wide fractions for blacks and whites who vote ($B^b$ and $B^w$) and who vote for

the Democrats ($\Lambda^b$ and $\Lambda^w$). These are each expressed as weighted averages of the precinct-level parameters:

$$B^b = \sum_{i=1}^{p} \frac{N_i^b \beta_i^b}{N^b}, \qquad B^w = \sum_{i=1}^{p} \frac{N_i^w \beta_i^w}{N^w}$$

$$\Lambda^b = \sum_{i=1}^{p} \frac{N_i^{bT} \lambda_i^b}{N^{bT}}, \qquad \Lambda^w = \sum_{i=1}^{p} \frac{N_i^{wT} \lambda_i^w}{N^{wT}} \tag{2.1}$$

where the number of blacks and whites of voting age in the entire district (i.e., in all $p$ precincts) are $N^b = \sum_{i=1}^{p} N_i^b$ and $N^w = \sum_{i=1}^{p} N_i^w$, respectively. These weighted averages do *not* equal the unweighted averages, except in the extremely unusual case where the black and white voting-age populations are identical within and across all precincts or, more generally, if the precinct parameters and the weights are independent. I also introduce notation for the unweighted average of $\beta_i^w$, in addition to that for $\beta_i^b$:

$$\mathfrak{B}^b = \frac{1}{p} \sum_{i=1}^{p} \beta_i^b, \qquad \mathfrak{B}^w = \frac{1}{p} \sum_{i=1}^{p} \beta_i^w \tag{2.2}$$

In general, we should be primarily interested in the precinct-level parameters ($\beta_i^b$, $\beta_i^w$, $\lambda_i^b$, and $\lambda_i^w$) in order to learn about geographic patterns in black and white turnout and voter support for each candidate and to extract the largest amount of information available from the ecological inference problem. These are the ultimate goals. However, it also makes sense to consider what district-wide summaries might be of interest. One possibility is the simple averages of the precinct-level parameters, $\mathfrak{B}^b$ and $\mathfrak{B}^w$ (and similarly for the $\lambda_i$'s), but these are of little substantive interest (even though they will sometimes prove convenient in the following chapters as intermediate results). Precincts are usually of very different sizes and have boundaries that are convenient rather than politically relevant. Instead, the aggregate values of these parameters for all people in the district ($B^b$ and $B^w$, as well as $\Lambda^b$, and $\Lambda^w$) are of considerable interest. In fact, the degree to which the average of the precinct parameters deviates from the population mean (that is, weighted average) is in part a result of the aggregation effects we would like to avoid.

Finally, I introduce $\theta_i^b$ (black vote for the Democratic candidate as proportion of the black voting-age population) and $\theta_i^w$ (white vote for

the Democrat as a proportion of the white voting-age population).

$$\theta_i^b = \frac{N_i^{bD}}{N_i^b}, \qquad \theta_i^w = \frac{N_i^{wD}}{N_i^w} \tag{2.3}$$

These parameters are of no intrinsic interest, but they will prove useful in intermediate stages for calculating some parameters of interest (since $\lambda_i^b = \theta_i^b/\beta_i^b$ and $\lambda_i^w = \theta_i^w/\beta_i^w$). The weighted averages of these parameters will also prove useful:

$$\Theta^b = \sum_{i=1}^{p} \frac{N_i^b \theta_i^b}{N^b}, \qquad \Theta^w = \sum_{i=1}^{p} \frac{N_i^w \theta_i^w}{N^w} \tag{2.4}$$