CAUSAL INFERENCE IN STATISTICS

A Primer

Judea Pearl
Madelyn Glymour
Nicholas P. Jewell

with website

WILEY

CAUSAL INFERENCE IN STATISTICS

CAUSAL INFERENCE IN STATISTICS

A PRIMER

Judea Pearl

Computer Science and Statistics, University of California, Los Angeles, USA

Madelyn Glymour

Philosophy, Carnegie Mellon University, Pittsburgh, USA

Nicholas P. Jewell

Biostatistics and Statistics, University of California, Berkeley, USA

WILEY

This edition first published 2016
© 2016 John Wiley & Sons Ltd

Registered office
John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, United Kingdom

For details of our global editorial offices, for customer services and for information about how to apply for permission to reuse the copyright material in this book please see our website at www.wiley.com.

The right of the author to be identified as the author of this work has been asserted in accordance with the Copyright, Designs and Patents Act 1988.

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, except as permitted by the UK Copyright, Designs and Patents Act 1988, without the prior permission of the publisher.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic books.

Designations used by companies to distinguish their products are often claimed as trademarks. All brand names and product names used in this book are trade names, service marks, trademarks or registered trademarks of their respective owners. The publisher is not associated with any product or vendor mentioned in this book.

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. It is sold on the understanding that the publisher is not engaged in rendering professional services and neither the publisher nor the author shall be liable for damages arising herefrom. If professional advice or other expert assistance is required, the services of a competent professional should be sought.

Library of Congress Cataloging-in-Publication Data applied for

ISBN: 9781119186847

A catalogue record for this book is available from the British Library.

Cover Image: © gmaydos/Getty

Typeset in 10/12pt TimesLTStd by SPi Global, Chennai, India

10 9 8 7 6 5 4 3 2

To my wife, Ruth, my greatest mentor.
– Judea Pearl

To my parents, who are the causes of me.
– Madelyn Glymour

To Debra and Britta, who inspire me every day.
– Nicholas P. Jewell

Contents

About the Authors **ix**

Preface **xi**

List of Figures **xv**

About the Companion Website **xix**

1 Preliminaries: Statistical and Causal Models **1**
1.1 Why Study Causation 1
1.2 Simpson's Paradox 1
1.3 Probability and Statistics 7
1.3.1 Variables 7
1.3.2 Events 8
1.3.3 Conditional Probability 8
1.3.4 Independence 10
1.3.5 Probability Distributions 11
1.3.6 The Law of Total Probability 11
1.3.7 Using Bayes' Rule 13
1.3.8 Expected Values 16
1.3.9 Variance and Covariance 17
1.3.10 Regression 20
1.3.11 Multiple Regression 22
1.4 Graphs 24
1.5 Structural Causal Models 26
1.5.1 Modeling Causal Assumptions 26
1.5.2 Product Decomposition 29

2 Graphical Models and Their Applications **35**
2.1 Connecting Models to Data 35
2.2 Chains and Forks 35
2.3 Colliders 40
2.4 *d*-separation 45
2.5 Model Testing and Causal Search 48

3 The Effects of Interventions 53
3.1 Interventions 53
3.2 The Adjustment Formula 55
3.2.1 To Adjust or not to Adjust? 58
3.2.2 Multiple Interventions and the Truncated Product Rule 60
3.3 The Backdoor Criterion 61
3.4 The Front-Door Criterion 66
3.5 Conditional Interventions and Covariate-Specific Effects 70
3.6 Inverse Probability Weighing 72
3.7 Mediation 75
3.8 Causal Inference in Linear Systems 78
3.8.1 Structural versus Regression Coefficients 80
3.8.2 The Causal Interpretation of Structural Coefficients 81
3.8.3 Identifying Structural Coefficients and Causal Effect 83
3.8.4 Mediation in Linear Systems 87

4 Counterfactuals and Their Applications 89
4.1 Counterfactuals 89
4.2 Defining and Computing Counterfactuals 91
4.2.1 The Structural Interpretation of Counterfactuals 91
4.2.2 The Fundamental Law of Counterfactuals 93
4.2.3 From Population Data to Individual Behavior – An Illustration 94
4.2.4 The Three Steps in Computing Counterfactuals 96
4.3 Nondeterministic Counterfactuals 98
4.3.1 Probabilities of Counterfactuals 98
4.3.2 The Graphical Representation of Counterfactuals 101
4.3.3 Counterfactuals in Experimental Settings 103
4.3.4 Counterfactuals in Linear Models 106
4.4 Practical Uses of Counterfactuals 107
4.4.1 Recruitment to a Program 107
4.4.2 Additive Interventions 109
4.4.3 Personal Decision Making 111
4.4.4 Sex Discrimination in Hiring 113
4.4.5 Mediation and Path-disabling Interventions 114
4.5 Mathematical Tool Kits for Attribution and Mediation 116
4.5.1 A Tool Kit for Attribution and Probabilities of Causation 116
4.5.2 A Tool Kit for Mediation 120

References 127

Index 133

About the Authors

Judea Pearl is Professor of Computer Science and Statistics at the University of California, Los Angeles, where he directs the Cognitive Systems Laboratory and conducts research in artificial intelligence, causal inference and philosophy of science. He is a Co-Founder and Editor of the *Journal of Causal Inference* and the author of three landmark books in inference-related areas. His latest book, *Causality: Models, Reasoning and Inference* (Cambridge, 2000, 2009), has introduced many of the methods used in modern causal analysis. It won the Lakatosh Award from the London School of Economics and is cited by more than 13,000 scientific publications.

Pearl is a member of the National Academy of Sciences, the National Academy of Engineering, and a Founding Fellow of the Association for Artificial Intelligence. He is a recipient of numerous prizes and awards, including the Technion's Harvey Prize and the ACM Alan Turing Award for fundamental contributions to probabilistic and causal reasoning.

Madelyn Glymour is a data analyst at Carnegie Mellon University, and a science writer and editor for the Cognitive Systems Laboratory at UCLA. Her interests lie in causal discovery and in the art of making complex concepts accessible to broad audiences.

Nicholas P. Jewell is Professor of Biostatistics and Statistics at the University of California, Berkeley. He has held various academic and administrative positions at Berkeley since his arrival in 1981, most notably serving as Vice Provost from 1994 to 2000. He has also held academic appointments at the University of Edinburgh, Oxford University, the London School of Hygiene and Tropical Medicine, and at the University of Kyoto. In 2007, he was a Fellow at the Rockefeller Foundation Bellagio Study Center in Italy.

Jewell is a Fellow of the American Statistical Association, the Institute of Mathematical Statistics, and the American Association for the Advancement of Science (AAAS). He is a past winner of the Snedecor Award and the Marvin Zelen Leadership Award in Statistical Science from Harvard University. He is currently the Editor of the *Journal of the American Statistical Association – Theory & Methods*, and Chair of the Statistics Section of AAAS. His research focuses on the application of statistical methods to infectious and chronic disease epidemiology, the assessment of drug safety, time-to-event analyses, and human rights.

Preface

When attempting to make sense of data, statisticians are invariably motivated by causal questions. For example, "How effective is a given treatment in preventing a disease?"; "Can one estimate obesity-related medical costs?"; "Could government actions have prevented the financial crisis of 2008?"; "Can hiring records prove an employer guilty of sex discrimination?"

The peculiar nature of these questions is that they cannot be answered, or even articulated, in the traditional language of statistics. In fact, only recently has science acquired a mathematical language we can use to express such questions, with accompanying tools to allow us to answer them from data.

The development of these tools has spawned a revolution in the way causality is treated in statistics and in many of its satellite disciplines, especially in the social and biomedical sciences. For example, in the technical program of the 2003 Joint Statistical Meeting in San Francisco, there were only 13 papers presented with the word "cause" or "causal" in their titles; the number of such papers exceeded 100 by the Boston meeting in 2014. These numbers represent a transformative shift of focus in statistics research, accompanied by unprecedented excitement about the new problems and challenges that are opening themselves to statistical analysis. Harvard's political science professor Gary King puts this revolution in historical perspective: "More has been learned about causal inference in the last few decades than the sum total of everything that had been learned about it in all prior recorded history."

Yet this excitement remains barely seen among statistics educators, and is essentially absent from statistics textbooks, especially at the introductory level. The reasons for this disparity is deeply rooted in the tradition of statistical education and in how most statisticians view the role of statistical inference.

In Ronald Fisher's influential manifesto, he pronounced that "the object of statistical methods is the reduction of data" (Fisher 1922). In keeping with that aim, the traditional task of making sense of data, often referred to generically as "inference," became that of finding a parsimonious mathematical description of the joint distribution of a set of variables of interest, or of specific parameters of such a distribution. This general strategy for inference is extremely familiar not just to statistical researchers and data scientists, but to anyone who has taken a basic course in statistics. In fact, many excellent introductory books describe smart and effective ways to extract the maximum amount of information possible from the available data. These books take the novice reader from experimental design to parameter estimation and hypothesis testing in great detail. Yet the aim of these techniques are invariably the

description of data, not of the process responsible for the data. Most statistics books do not even have the word "causal" or "causation" in the index.

Yet the fundamental question at the core of a great deal of statistical inference is causal; do changes in one variable cause changes in another, and if so, how much change do they cause? In avoiding these questions, introductory treatments of statistical inference often fail even to discuss whether the parameters that are being estimated are the relevant quantities to assess when interest lies in cause and effects.

The best that most introductory textbooks do is this: First, state the often-quoted aphorism that "association does not imply causation," give a short explanation of confounding and how "lurking variables" can lead to a misinterpretation of an apparent relationship between two variables of interest. Further, the boldest of those texts pose the principal question: "How can a causal link between x and y be established?" and answer it with the long-standing "gold standard" approach of resorting to randomized experiment, an approach that to this day remains the cornerstone of the drug approval process in the United States and elsewhere.

However, given that most causal questions cannot be addressed through random experimentation, students and instructors are left to wonder if there is anything that can be said with any reasonable confidence in the absence of pure randomness.

In short, by avoiding discussion of causal models and causal parameters, introductory textbooks provide readers with no basis for understanding how statistical techniques address scientific questions of causality.

It is the intent of this primer to fill this gnawing gap and to assist teachers and students of elementary statistics in tackling the causal questions that surround almost any nonexperimental study in the natural and social sciences. We focus here on simple and natural methods to define *causal* parameters that we wish to understand and to show what assumptions are necessary for us to estimate these parameters in observational studies. We also show that these assumptions can be expressed mathematically and transparently and that simple mathematical machinery is available for translating these assumptions into estimable causal quantities, such as the effects of treatments and policy interventions, to identify their testable implications.

Our goal stops there for the moment; we do not address in any detail the optimal parameter estimation procedures that use the data to produce effective statistical estimates and their associated levels of uncertainty. However, those ideas—some of which are relatively advanced—are covered extensively in the growing literature on causal inference. We thus hope that this short text can be used in conjunction with standard introductory statistics textbooks like the ones we have described to show how statistical models and inference can easily go hand in hand with a thorough understanding of causation.

It is our strong belief that if one wants to move beyond mere description, statistical inference cannot be effectively carried out without thinking carefully about causal questions, and without leveraging the simple yet powerful tools that modern analysis has developed to answer such questions. It is also our experience that thinking causally leads to a much more exciting and satisfying approach to both the simplest and most complex statistical data analyses. This is not a new observation. Virgil said it much more succinctly than we in 29 BC:

"Felix, qui potuit rerum cognoscere causas" (Virgil 29 BC)
(Lucky is he who has been able to understand the causes of things)

The book is organized in four chapters.

Chapter 1 provides the basic statistical, probabilistic, and graphical concepts that readers will need to understand the rest of the book. It also introduces the fundamental concepts of causality, including the causal model, and explains through examples how the model can convey information that pure data are unable to provide.

Chapter 2 explains how causal models are reflected in data, through patterns of statistical dependencies. It explains how to determine whether a data set complies with a given causal model, and briefly discusses how one might search for models that explain a given data set.

Chapter 3 is concerned with how to make predictions using causal models, with a particular emphasis on predicting the outcome of a policy intervention. Here we introduce techniques of reducing confounding bias using adjustment for covariates, as well as inverse probability weighing. This chapter also covers mediation analysis and contains an in-depth look at how the causal methods discussed thus far work in a linear system. Key to these methods is the fundamental distinction between regression coefficients and structural parameters, and how students should use both to predict causal effects in linear models.

Chapter 4 introduces the concept of counterfactuals—what would have happened, had we chosen differently at a point in the past—and discusses how we can compute them, estimate their probabilities, and what practical questions we can answer using them. This chapter is somewhat advanced, compared to its predecessors, primarily due to the novelty of the notation and the hypothetical nature of the questions asked. However, the fact that we read and compute counterfactuals using the same scientific models that we used in previous chapters should make their analysis an easy journey for students and instructors. Those wishing to understand counterfactuals on a friendly mathematical level should find this chapter a good starting point, and a solid basis for bridging the model-based approach taken in this book with the potential outcome framework that some experimentalists are pursuing in statistics.

Acknowledgments

This book is an outgrowth of a graduate course on causal inference that the first author has been teaching at UCLA in the past 20 years. It owes many of its tools and examples to former members of the Cognitive Systems Laboratory who participated in the development of this material, both as researchers and as teaching assistants. These include Alex Balke, David Chickering, David Galles, Dan Geiger, Moises Goldszmidt, Jin Kim, George Rebane, Ilya Shpitser, Jin Tian, and Thomas Verma.

We are indebted to many colleagues from whom we have learned much about causal problems, their solutions, and how to present them to general audiences. These include Clark and Maria Glymour, for providing patient ears and sound advice on matters of both causation and writing, Felix Elwert and Tyler VanderWeele for insightful comments on an earlier version of the manuscript, and the many visitors and discussants to the UCLA Causality blog who kept the discussion lively, occasionally controversial, but never boring (causality.cs.ucla.edu/blog).

Elias Bareinboim, Bryant Chen, Andrew Forney, Ang Li, Karthika Mohan, reviewed the text for accuracy and transparency. Ang and Andrew also wrote solutions to the study questions, which are available to instructors from the publisher, see <http://bayes.cs.ucla.edu/PRIMER/CIS-Manual-PUBLIC.pdf>.

The manuscript was most diligently typed, processed, illustrated, and proofed by Kaoru Mulvihill at UCLA. Debbie Jupe and Heather Kay at Wiley deserve much credit for recognizing and convincing us that a book of this scope is badly needed in the field, and for encouraging us throughout the production process.

Finally, the National Science Foundation and the Office of Naval Research deserve acknowledgment for faithfully and consistently sponsoring the research that led to these results, with special thanks to Behzad Kamgar-Parsi.