Springer Series in Statistics

Paul R. Rosenbaum

Design of Observational Studies

Second Edition

Springer

Springer Series in Statistics

Series Editors:
Peter Bühlmann, Peter Diggle, Ursula Gather, Scott Zeger

Past Editors:
Peter Bickel, Ingram Olkin, Nanny Wermuth

Founding Editors:
David Brillinger, Stephen Fienberg, Joseph Gani, John Hartigan, Jack Kiefer, Klaus Krickeberg

Springer Series in Statistics (SSS) is a series of monographs of general interest that discuss statistical theory and applications.

The series editors are currently Peter Bühlmann, Peter Diggle, Ursula Gather, and Scott Zeger. Peter Bickel, Ingram Olkin, and Stephen Fienberg were editors of the series for many years.

More information about this series at http://www.springer.com/series/692

Paul R. Rosenbaum

Design of Observational Studies

Second Edition

Springer

Paul R. Rosenbaum
Statistics Department
Wharton School
University of Pennsylvania
Philadelphia, PA, USA

ISSN 0172-7397 ISSN 2197-568X (electronic)
Springer Series in Statistics
ISBN 978-3-030-46404-2 ISBN 978-3-030-46405-9 (eBook)
https://doi.org/10.1007/978-3-030-46405-9

© Springer Nature Switzerland AG 2010, 2020
This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Simplicity of form is not necessarily simplicity of experience.

Robert Morris, writing about art.

Simplicity is not a given. It is an achievement.

William H. Gass, writing about literature.

Simplicity . . . is a very important matter which must be constantly borne in mind.

Sir David Cox, writing about experiments.

For Judy

Preface to the First Edition

An observational study is an empiric investigation of effects caused by treatments when randomized experimentation is unethical or infeasible. The quality and strength of evidence provided by an observational study is determined largely by its design. Excellent methods of analysis will not salvage a poorly designed study.

The line between design and analysis is easier to draw in practice than it is in theory. In practice, the design of an observational study consists of all activities that precede the examination or use of those outcome measures that will be the basis for the study's conclusions. Unlike experiments, in some observational studies, the outcomes may exist as measurements prior to the design of the study; it is their examination and use, not their existence, that separates design from analysis. Aspects of design include the framing of scientific questions to permit empirical investigation, the choice of a context in which to conduct the investigation, decisions about what data to collect, where and how to collect it, matching to remove bias from measured covariates, strategies and tactics to limit uncertainty caused by covariates not measured, and sample splitting to guide design using individuals who will not be included in the final analysis. In practice, design ends and analysis begins when outcomes are examined for individuals who will be the basis of the study's conclusions. An observational study that begins by examining outcomes is a formless, undisciplined investigation that lacks design.

In theory, design anticipates analysis. The analysis is ever present in design, as any goal is ever present in any organized effort, as a goal is necessary to organize effort. One seeks to ask questions and collect data so that results will be decisive when analyzed. To end well, how should we begin?

Philadelphia, PA, USA — Paul R. Rosenbaum
5 August 2009

Preface to the Second Edition

The second edition differs from the first in two ways: there is additional material, and there is a new relationship with the `R` statistical language, including a companion `R` package `DOS2`. Also, a few errors have been corrected.

New Topics in the Second Edition

There are four new chapters:

- Chapter 7: Some Counterclaims Undermine Themselves
- Chapter 19: Choice of Test Statistic
- Chapter 20: Evidence Factors
- Chapter 21: Constructing Several Comparison Groups

There are numerous additions and revisions to existing chapters, including the following new sections:

- Section 2.4.4: The Average Treatment Effect
- Section 3.6: Amplification of a Sensitivity Analysis
- Section 5.4: Strengthening Weak Instruments
- Several new sections in Chap. 10
- Several new sections in Chap. 11
- Section 13.6: Isolation in Natural Experiments
- Section 18.3: Can Coherence Be Discovered?

R in the Second Edition

There is a new R package DOS2 available at cran. It contains software and data, and it reproduces many analyses from the second edition. There is also a more limited, new package DOS for the first edition. In addition to different contents, the help-files for DOS and DOS2 refer to the first and second editions, respectively.

The first edition contained the text of some R functions. These have been removed and placed in the packages DOS2 and DOS.

Since the publication of the first edition, there have been many new R packages devoted to the design and analysis of observational studies. The second edition describes some of these packages. Specifically, several chapters of the second edition end with sections entitled "Software" or "Data" describing available resources in R. The index contains entries for "R packages," "software," and "data."

A shinyapp at http://rosenbap.shinyapps.io/learnShiny/ permits you to conduct an interactive sensitivity analysis for an observational study of smoking and periodontal disease. No knowledge of R is required. The shinyapp runs R in the background, but use of the app is point-and-click. The app illustrates several issues: (1) randomization inference using Huber's m-statistics (Sect. 2.8), (2) sensitivity analysis for P-values, estimates, and confidence intervals (Sect. 3.4), (3) amplification of sensitivity analysis (Sect. 3.6), and (4) the fact that a poor choice of test statistic leads to an understatement of insensitivity to unmeasured bias (Chap. 19). The app will show the R code if you request it.

Philadelphia, PA, USA Paul R. Rosenbaum
25 November 2019

Acknowledgments

I am in debt to many people: to Jeffrey Silber, Dylan Small, Mike Baiocchi, Joseph Gastwirth, Robert Greevy, Ben Hansen, Ruth Heller, Guido Imbens, Bikram Karmakar, Bo Lu, Samuel Pimentel, Ruoqi Yu, and José Zubizarreta for recent collaborations I describe here in detail; to my teacher, adviser, and coauthor Donald Rubin, from whom I learned a great deal; to colleagues, coauthors, or former students Katrina Armstrong, Susan Bakewell-Sachs, Lisa Bellini, T. Behringer, Shawna Calhoun, Jing Cheng, Magdalena Cerdá, Avital Cnaan, Shoshana Daniel, Ashkan Ertefaie, Gabriel Escobar, Orit Even-Shoshan, Lee Fleisher, Kevin Fox, Bruce Giantonio, Neera Goyal, Sam Gu, Elizabeth Halloran, Amelia Haviland, Alexander Hill, Lauren Hochman, Robert Hornik, Jesse Hsu, Kamal Itani, Rachel Kelz, Abba Krieger, Marshall Joffe, Kwonsang Lee, Yunfei Paul Li, Scott Lorch, Rossi X. Luo, Justin Ludwig, Barbara Medoff-Cooper, Lanyu Mi, Andrea Millman, Kewei Ming, Dan Nagin, Mark Neuman, Bijan Niknam, Ricardo Paredes, Dan Polsky, Kate Propert, Tom Randall, Caroline Reinke, Joseph Reiter, Patrick Romano, Amy Rosen, Richard Ross, Sandy Schwartz, Morgan Sellers, James Sharpe, Tom Ten Have, Richard Tremblay, Kevin Volpp, Yanli Wang, Wei Wang, Dan Yang, Fan Yang, Frank Yoon, Elaine Zanutto, and Qingyuan Zhao for collaborations I describe more briefly; to Judith McDonald, Keith Goldfeld, Bikram Karmakar, Sue Marcus, John Newell, Luke Keele, Dylan Small, José Zubizarreta, and anonymous reviewers for comments on the book; to Joshua Angrist, David Card, Susan Dynarski, Alan Krueger, and Victor Lavy for making micro-data from their research available in one form or another.

Parts of the first edition were written while I was on sabbatical. The hospitality of the Department of Statistics at Columbia University and the Department of Economics and the Statistics Program at the National University of Ireland at Galway is gratefully acknowledged. The work was supported in part by the Methodology, Measurement, and Statistics Program of the U.S. National Science Foundation.

Of course, my greatest debts are to Judy, Sarah, Hannah, Aaron, Dac, Mark, Josh, Paul, and Noah.

Contents

Part I Beginnings

1 Dilemmas and Craftsmanship 3
1.1 Those Confounded Vitamins 3
1.2 Cochran's Basic Advice 4
1.2.1 Treatments, Covariates, Outcomes 5
1.2.2 How were Treatments Assigned? 5
1.2.3 Were Treated and Control Groups Comparable? 5
1.2.4 Eliminating Plausible Alternatives to Treatment Effects 6
1.2.5 Exclusion Criteria 6
1.2.6 Exiting a Treatment Group After Treatment Assignment 7
1.2.7 Study Protocol 7
1.3 Maimonides' Rule 7
1.4 Seat Belts in Car Crashes 9
1.5 Money for College 10
1.6 Nature's "Natural Experiment" 12
1.7 What This Book Is About 13
1.7.1 Basic Structure 13
1.7.2 Structure of Part I: Beginnings 14
1.7.3 Structure of Part II: Matching 15
1.7.4 Structure of Part III: Design Sensitivity 15
1.7.5 Structure of Part IV: Enhanced Design 15
1.7.6 Structure of Part V: Planning Analysis 16
1.7.7 A Less Technical Introduction to Observational Studies 16
1.7.8 Dependence Among Chapters 17
1.8 Further Reading 18
Bibliography 20

2 Causal Inference in Randomized Experiments 23
2.1 Two Versions of the National Supported Work Experiment 23
2.1.1 A Version with 185 Pairs and a Version with 5 Pairs 23
2.1.2 Basic Notation 25
2.2 Treatment Effects in Randomized Experiments 28
2.2.1 Potential Responses Under Alternative Treatments 28
2.2.2 Covariates and Outcomes 29
2.2.3 Possible Treatment Assignments and Randomization 29
2.2.4 Interference Between Units 31
2.3 Testing the Null Hypothesis of No Treatment Effect 32
2.3.1 Treated–Control Differences When the Null Hypothesis Is True 32
2.3.2 The Randomization Distribution of the Mean Difference 33
2.3.3 The Randomization Distribution of Wilcoxon's Statistic 39
2.4 Testing Other Hypotheses; Confidence Intervals; Point Estimates 43
2.4.1 Testing a Constant, Additive Treatment Effect 43
2.4.2 Confidence Intervals for a Constant, Additive Effect 45
2.4.3 Hodges–Lehmann Point Estimates of Effect 47
2.4.4 The Average Treatment Effect 48
2.4.5 Testing General Hypotheses About Treatment Effects 50
2.4.6 Multiplicative Effects; Tobit Effects 52
2.5 Attributable Effects 54
2.5.1 Why Use Attributable Effects? 54
2.5.2 Aligned Responses: Shifting Attention from Pairs to Individuals 55
2.5.3 Thinking About Heterogeneous Effects by Focusing on Small Groups of Pairs 56
2.5.4 Computations; Null Distribution 58
2.6 Internal and External Validity 62
2.7 Summary 63
2.8 Appendix: Randomization Distribution of m-Statistics 64
2.8.1 Giving Observations Controlled Weight Using a ψ Function 64
2.8.2 Scaling 66
2.8.3 Randomization Test of a Hypothesized Additive Effect, $H_0 : r_{Tij} = r_{Cij} + \tau_0$ 66
2.8.4 m-Tests in the NSW Experiment 67
2.9 Further Reading 67

2.10 Software 68
2.11 Data 68
Bibliography 69

3 Two Simple Models for Observational Studies 73
3.1 The Population Before Matching 73
3.2 The Ideal Matching 74
3.3 A Naïve Model: People Who Look Comparable Are Comparable 78
3.3.1 Assigning Treatments by Flipping Biased Coins Whose Unknown Biases Are Determined by Observed Covariates 78
3.3.2 Balancing Property of the Propensity Score 81
3.3.3 Propensity Scores and Ignorable Treatment Assignment 83
3.3.4 Summary: Separating Two Tasks, One Mechanical, the Other Scientific 83
3.4 Sensitivity Analysis: People Who Look Comparable May Differ 84
3.4.1 What Is Sensitivity Analysis? 84
3.4.2 The Sensitivity Analysis Model: Quantitative Deviation from Random Assignment 85
3.4.3 Sensitivity Analysis Model When Pairs Are Matched for Observed Covariates 87
3.5 Welding Fumes and DNA Damage 88
3.5.1 Sensitivity Analysis When Testing the Hypothesis of No Treatment Effect 88
3.5.2 Computations 91
3.5.3 Sensitivity Analysis for a Confidence Interval 93
3.5.4 Sensitivity Analysis for Point Estimates 94
3.6 Amplification of a Sensitivity Analysis 95
3.6.1 What Is an Amplification? 95
3.6.2 Sensitivity Analysis in Terms of (Λ, Δ) 96
3.6.3 Precise Meaning of (Λ, Δ) 97
3.7 Bias Due to Incomplete Matching 99
3.8 Summary 100
3.9 Further Reading 101
3.10 Software 102
3.11 Data 102
Bibliography 103

4 Competing Theories Structure Design 111
4.1 How Stones Fall 111
4.2 The Permanent-Debt Hypothesis 114
4.3 Guns and Misdemeanors 116
4.4 The Dutch Famine of 1944–1945 117

4.5 Replicating Effects and Biases 118
4.6 Reasons for Effects 121
4.7 The Drive for System 124
4.8 Further Reading 126
Bibliography 126

5 Opportunities, Devices, and Instruments 129
5.1 Opportunities 129
5.1.1 The Well-Ordered World 129
5.1.2 Questions 130
5.1.3 Solutions 130
5.2 Devices 132
5.2.1 Disambiguation 132
5.2.2 Multiple Control Groups 132
5.2.3 Coherence Among Several Outcomes 134
5.2.4 Known Effects 138
5.2.5 Doses of Treatment 141
5.2.6 Differential Effects and Generic Biases 145
5.3 Instruments 148
5.3.1 What Is an Instrument? 148
5.3.2 Example: Noncompliance in a Double-Blind Randomized Trial 149
5.3.3 Example: Maimonides' Rule 150
5.3.4 Notation for an Instrument in a Paired Encouragement Design 150
5.3.5 The Hypothesis that Effect Is Proportional to Dose 152
5.3.6 Inference About β 152
5.3.7 Example: IV Analysis for Maimonides' Rule 154
5.3.8 Effect Ratios 156
5.3.9 Is the Validity of an Instrument Testable? 158
5.3.10 When and Why Are Instruments Valuable? 158
5.4 Strengthening Weak Instruments 159
5.4.1 Why Strengthen an Instrument? 159
5.4.2 A Popular Instrument in Health Outcomes Research: Distance to a Hospital 160
5.4.3 How Is an Instrument Strengthened? 161
5.4.4 Use of Matching in Strengthening an Instrument 163
5.5 Summary 164
5.6 Further Reading 165
5.7 Software 165
5.8 Data 166
Bibliography 166

6 Transparency 175
References 177

7 Some Counterclaims Undermine Themselves 179
7.1 Appraising Counterclaims 179
7.1.1 Types of Counterclaims 179
7.1.2 The Logic of Counterclaims That Undermine Themselves 180
7.2 An Example: Safety Belts, Injury, and Ejection 181
7.2.1 A First Look at an Example: A Claim Prior to a Counterclaim 181
7.2.2 The Counterclaim of Selection Bias and Secondary Outcomes 183
7.2.3 A Second Look at an Example in Light of a Counterclaim 184
7.3 Discussion 185
7.3.1 Anticipating Counterclaims 185
7.3.2 Some Theory 186
7.4 Further Reading 186
7.5 Data 186
Bibliography 186

Part II Matching

8 A Matched Observational Study 191
8.1 Is More Chemotherapy More Effective? 191
8.2 Matching for Observed Covariates 192
8.3 Outcomes in Matched Pairs 195
8.4 Summary 197
8.5 Further Reading 199
References 199

9 Basic Tools of Multivariate Matching 201
9.1 A Small Example 201
9.2 Propensity Score 204
9.3 Distance Matrices 205
9.4 Optimal Pair Matching 211
9.5 Optimal Matching with Multiple Controls 213
9.6 Optimal Full Matching 218
9.7 Efficiency 223
9.8 Summary 224
9.9 Further Reading 224
9.10 Software 225
9.11 Data 225
Bibliography 226

10 Various Practical Issues in Matching 229
10.1 Checking Covariate Balance 229
10.2 Simulated Randomized Experiments as Diagnostics 232

10.3 Near-Exact Matching 234
10.4 Exact Matching 236
10.5 Directional Penalties 237
10.6 Missing Covariate Values 238
10.7 Networks and Sparse Representations of Matching 239
10.8 Constraints on Individual Distances 240
10.9 Clustered Treatment Assignment 241
10.10 Further Reading 242
10.11 Software 242
Bibliography 242

11 Fine Balance 245
11.1 What Is Fine Balance? 245
11.2 Constructing a Finely Balanced Control Group 246
11.3 Controlling Imbalance When Fine Balance Is Not Feasible 249
11.4 Fine Balance, Exact Matching and Near-Exact Matching 252
11.5 Near-Fine Balance 253
11.6 Refined Balance 254
11.7 Strength K Balance 255
11.8 Cardinality Matching 255
11.8.1 What Is Cardinality Matching? 255
11.8.2 Cardinality Matching and Outcome Heterogeneity 256
11.8.3 Cardinality Matching and Effect Modification 257
11.9 Further Reading 257
11.10 Software 257
11.11 Data 258
Bibliography 258

12 Matching Without Groups 261
12.1 Matching Without Groups: Nonbipartite Matching 261
12.1.1 What Is Nonbipartite Matching? 261
12.1.2 Treatment-Control Matching Using an Algorithm for Nonbipartite Matching 262
12.1.3 Matching with Doses 263
12.1.4 Matching with Several Groups 264
12.2 Some Practical Aspects of Matching Without Groups 265
12.2.1 An Odd Number of Subjects 265
12.2.2 Discarding Some Subjects 266
12.2.3 Balanced Incomplete Block Designs with Three Groups 266
12.2.4 Propensity Scores for Several Groups 267
12.3 Matching with Doses and Two Control Groups 267
12.3.1 Does the Minimum Wage Reduce Employment? 267
12.3.2 Optimal Matching to Form Two Independent Comparisons 268

12.3.3 Difference in Change in Employment with Two Control Groups ... 271
12.4 Further Reading ... 274
12.5 Software ... 275
12.6 Data ... 275
Bibliography ... 275

13 Risk-Set Matching ... 279
13.1 Does Cardiac Transplantation Prolong Life? ... 279
13.2 Risk-Set Matching in a Study of Surgery for Interstitial Cystitis ... 280
13.3 Maturity at Discharge from a Neonatal Intensive Care Unit ... 284
13.4 Joining a Gang at Age 14 ... 287
13.5 Some Theory ... 288
13.6 Isolation in Natural Experiments ... 290
13.6.1 Brief Review of Differential Effects and Generic Biases ... 290
13.6.2 What Is Isolation? ... 290
13.6.3 Twins Versus a Single Birth, and Their Impact on Labor Supply ... 291
13.6.4 Mass and Safety in Vehicle Crashes Resulting in a Fatality ... 292
13.7 Further Reading ... 294
13.8 Software ... 295
Bibliography ... 295

14 Matching in R ... 299
14.1 Optimal Matching Using R ... 299
14.2 Data ... 300
14.3 Propensity Score ... 302
14.4 Covariates with Missing Values ... 303
14.5 Distance Matrix ... 304
14.6 Constructing the Match ... 305
14.7 Checking Covariate Balance ... 308
14.8 College Outcomes ... 308
14.9 Further Reading ... 310
14.10 Software ... 311
14.11 Data ... 312
Bibliography ... 312

Part III Design Sensitivity

15 The Power of a Sensitivity Analysis and Its Limit ... 317
15.1 The Power of a Test in a Randomized Experiment ... 317
15.1.1 What Is the Power of a Test? ... 317
15.1.2 A Pep Talk About Statistical Power ... 319

15.1.3 Computing Power in a Randomized Experiment: The Two Steps 321
15.1.4 Step 1: Determining the Critical Value Assuming the Null Hypothesis Is True 321
15.1.5 Step 2: Determining the Power Assuming the Null Hypothesis Is False 323
15.1.6 A Simple Case: Constant Effect with Random Errors 324
15.2 Power of a Sensitivity Analysis in an Observational Study 326
15.2.1 What Is the Power of a Sensitivity Analysis? 326
15.2.2 Computing the Power of a Sensitivity Analysis: The Two Steps 327
15.2.3 Step 2: Determining the Power When the Null Hypothesis Is False and There Is No Unobserved Bias 328
15.2.4 A First Look At the Power of a Sensitivity Analysis 329
15.3 Design Sensitivity 330
15.3.1 A First Look at Design Sensitivity 330
15.3.2 A Formula for Design Sensitivity 331
15.3.3 Computing Design Sensitivity with Additive Effects and iid Errors 332
15.4 Summary 333
15.5 Further Reading 333
Bibliography 335

16 Heterogeneity and Causality 337
16.1 J.S. Mill and R.A. Fisher: Reducing Heterogeneity or Introducing Random Assignment 337
16.2 A Larger, More Heterogeneous Study Versus a Smaller, Less Heterogeneous Study 339
16.2.1 Large I or Small σ: Which is Better? 339
16.2.2 A Simulated Example 340
16.2.3 Power Comparisons with Normal, Logistic, and Cauchy Errors 342
16.2.4 Design Sensitivity 342
16.3 Heterogeneity and the Sensitivity of Point Estimates 343
16.4 Examples of Efforts to Reduce Heterogeneity 344
16.4.1 Twins 344
16.4.2 Road Hazards 345
16.4.3 The Genetically Engineered Mice of Microeconomics 345
16.4.4 Motorcycle Helmets 346
16.5 Summary 346
16.6 Further Reading 347
Bibliography 347

17 Uncommon but Dramatic Responses to Treatment 349
17.1 Large Effects, Now and Then 349
17.1.1 Are Large but Rare Effects Insensitive to Unmeasured Biases? 349
17.1.2 Review of Sect. 2.5: Measuring Large but Uncommon Effects 351
17.2 Two Examples 352
17.2.1 Chemotherapy Intensity and Toxicity in Ovarian Cancer 352
17.2.2 DNA Adducts Among Aluminum Production Workers 353
17.3 Properties of a Paired Version of Salsburg's Model 354
17.4 Design Sensitivity for Uncommon but Dramatic Effects 357
17.4.1 Design Sensitivity of Stephenson's Test 357
17.4.2 Design Sensitivity of Stephenson's Test under Salsburg's Model 358
17.5 Summary 359
17.6 Further Reading 359
17.7 Software 360
17.8 Data 360
Bibliography 361

18 Anticipated and Discovered Patterns of Response 363
18.1 Using Design Sensitivity to Evaluate Devices 363
18.2 Coherence 364
18.2.1 Notation with Several Responses 364
18.2.2 Responses with a Multivariate Normal Distribution 365
18.2.3 Numerical Results for Bivariate Normal Responses 366
18.2.4 Practical Implementation for General $\boldsymbol{\lambda}$ 369
18.3 Can Coherence Be Discovered? 371
18.3.1 Using Split Samples to Plan for Coherence 371
18.3.2 Considering Every Possible $\boldsymbol{\lambda}$ 371
18.3.3 Hedged Bets About $\boldsymbol{\lambda}$ 373
18.3.4 Summary 374
18.4 Doses 374
18.4.1 Another Way to Write a Signed Rank Statistic 375
18.4.2 The Favorable Situation with Doses 376
18.4.3 A Formula for the Design Sensitivity with Doses 377
18.4.4 Numerical Evaluation of the Design Sensitivity 378
18.5 Example: Maimonides' Rule 379
18.6 Reactive Doses 380
18.7 Further Reading 383
18.8 Software 383
18.9 Data 383
Bibliography 384

19 Choice of Test Statistic 387
19.1 Choice of Test Statistic Affects Design Sensitivity 387
19.1.1 Design Anticipates Analysis 387
19.1.2 A Simple Case: Step-Rank Statistics, Additive Effects, Normal Errors 388
19.2 Statistics Built for Superior Design Sensitivity 390
19.2.1 Building New Statistics for Use in Sensitivity Analyses 390
19.2.2 A New U-Statistic with Superior Design Sensitivity 391
19.2.3 Example: Chemotherapy Associated Toxicity 392
19.2.4 Example: Lead in the Blood of Smokers 393
19.2.5 m-Statistics 393
19.2.6 Summary 394
19.3 Adaptive Inference 395
19.3.1 Using the Data to Choose a Test Statistic 395
19.3.2 Example: Adaptive Inference for Chemotherapy Associated Toxicity 397
19.3.3 Example: Adaptive Inference for Lead in the Blood of Smokers 397
19.4 Design Sensitivity and Dose–Response 398
19.4.1 Dose–Response and Evidence of Causality 398
19.4.2 Example: Smoking and Periodontal Disease 399
19.4.3 Ignoring Doses: Pair Differences in Periodontal Disease 400
19.4.4 The Crosscut Test 400
19.4.5 Design Sensitivity of the Crosscut Test 401
19.4.6 The Stratified Crosscut Test 402
19.4.7 An Adaptive Crosscut Test 403
19.4.8 The Crosscut Test and Evidence Factors 405
19.5 Bahadur Efficiency of Sensitivity Analyses 405
19.5.1 Where Do Things Stand? 405
19.5.2 Several Types of Efficiency 406
19.5.3 Using Bahadur Efficiency in Sensitivity Analyses 408
19.6 Further Reading 409
19.7 Software 409
19.8 Data 410
Bibliography 410

Part IV Enhanced Design

20 Evidence Factors 415
20.1 What Are Evidence Factors? 415
20.1.1 A Replication Should Disrupt a Plausible Bias 415
20.1.2 Replication and Evidence Factors 416
20.1.3 Example: Smoking and Periodontal Disease 416

20.1.4 Issues Developed in This Chapter 417
20.2 The Simplest Nontrivial Case: Renyi's Partial Ranks 418
20.2.1 DNA Damage Among Tannery Workers 418
20.2.2 Renyi's Partial Ranks in a Randomized Experiment Under the Null Hypothesis 419
20.2.3 Wilcoxon's Stratified Rank Sum Test 421
20.2.4 Two Evidence Factors Using Wilcoxon's Rank Sum Test 422
20.2.5 Limitations of Partial Ranks 424
20.2.6 Sensitivity Analyses for Evidence Factors 425
20.3 A Second Example: Smoking and Periodontal Disease 426
20.4 Appendix: Some Theory 427
20.4.1 A Tiny Example 427
20.4.2 Elementary Theory of Groups of Permutation Matrices 430
20.4.3 Probability Distributions on a Group of Permutation Matrices 432
20.4.4 Invariant Test Statistics 433
20.4.5 Ignoring One Evidence Factor 433
20.4.6 Symmetric Sensitivity Analyses for Biases That May Be Asymmetric 435
20.4.7 Fixing One Evidence Factor 437
20.4.8 Joint Behavior of Two Bounds on P-Values 438
20.4.9 Joint Distribution of P-Value Bounds for Two Evidence Factors 439
20.4.10 Another Example: Blocks of Size Three 440
20.4.11 Untidy Blocks 441
20.5 Further Reading 442
20.6 Software 442
20.7 Data 443
Bibliography 443

21 Constructing Several Comparison Groups 445
21.1 Why Compare Several Groups? 445
21.1.1 Uses of Several Comparison Groups 445
21.1.2 Issues When Constructing More Than One Comparison Group 446
21.2 Overlapping Comparison Groups and the Exterior Match 446
21.2.1 Two Entwined Comparison Groups 446
21.2.2 The Exterior Match 447
21.3 Optimal Tapered Matching 449
21.4 Building Matched Sets Using an Approximation Algorithm 450
21.4.1 A Near Optimal Solution to a Difficult Problem 450
21.4.2 What Is Near-Fine Balance with Three Groups? 451
21.4.3 An Approximation Algorithm 452

21.5 Is It Possible to Attenuate Unmeasured Bias? 453
21.5.1 Attenuation: Logically Possible But Small in Magnitude 453
21.5.2 Look More, Speculate Less 454
21.5.3 Smoking and Elevated Homocysteine Levels 455
21.6 Further Reading 457
21.7 Software 457
Bibliography 457

Part V Planning Analysis

22 After Matching, Before Analysis 461
22.1 Split Samples and Design Sensitivity 461
22.2 Are Analytic Adjustments Feasible? 463
22.2.1 Tapered Matching and the Exterior Match 467
22.3 Matching and Thick Description 467
22.3.1 Thick Description 467
22.3.2 What Is Thick Description? 468
22.3.3 Example: Mortality After Surgery 469
22.4 Further Reading 470
22.5 Software 471
References 471

23 Planning the Analysis 475
23.1 Plans Enable 475
23.2 Elaborate Theories 477
23.2.1 R.A. Fisher's Advice 477
23.2.2 What Should a Planned Analysis Accomplish? 477
23.3 Three Simple Plans with Two Control Groups 478
23.3.1 The Simplest Plan for Analysis with Two Control Groups 478
23.3.2 A Symmetric Plan for Analysis with Two Control Groups 481
23.3.3 Are the Two Control Groups Nearly Equivalent? 482
23.3.4 An Elementary Plan for Analysis with Two Control Groups 483
23.3.5 An Alternative Plan for Analysis with Two Control Groups 485
23.3.6 Summary 487
23.4 Sensitivity Analysis for Two Outcomes and Coherence 488
23.5 Sensitivity Analysis for Tests of Equivalence 489
23.6 Sensitivity Analysis for Equivalence and Difference 492
23.7 Summary 494
23.8 Further Reading 494
Appendix: Testing Hypotheses in Order 495
Bibliography 499

Summary: Key Elements of Design 503

Solutions to Common Problems 505
References 508

Symbols 511

Acronyms 515

Glossary of Statistical Terms 517

Further Reading 523
References 523

Suggested Readings for a Course 527
References 527

Index 529

Part I
Beginnings

Chapter 1
Dilemmas and Craftsmanship

Abstract This introductory chapter mentions some of the issues that arise in observational studies and describes a few well designed studies. Section 1.7 outlines the book, describes its structure, and suggests alternative ways to read it.

1.1 Those Confounded Vitamins

On 22 May 2004, the *Lancet* published two articles, one entitled "When are observational studies as credible as randomized trials?" by Jan Vandenbroucke [63], the other entitled "Those confounded vitamins: What can we learn from the differences between observational versus randomized trial evidence?" by Debbie Lawlor, George Smith, Richard Bruckdorfer, Devi Kundu, and Shah Ebrahim [34].

In a randomized experiment or trial, a coin is flipped to decide whether the next person is assigned to treatment or control, whereas in an observational study, treatment assignment is not under experimental control. Despite the optimism of the first title and the pessimism of the second, both articles struck a balance, perhaps with a slight tilt towards pessimism. Vandenbroucke reproduced one of Jim Borgman's political cartoons in which a TV newsman sits below both a banner reading "Today's Random Medical News" and three spinners which have decided that "coffee can cause depression in twins." Deadpan, the newsman says, "According to a report released today. . . ." The cartoon reappeared in a recent report of the Academy of Medical Sciences that discusses observational studies in some detail [52, page 19].

Lawlor et al. begin by noting that a large observational study published in the *Lancet* [32] had found a strong, statistically significant negative association between coronary heart disease mortality and level of vitamin C in blood, having used a model to adjust for other variables such as age, blood pressure, diabetes, and smoking. Adjustments using a model attempt to compare people who are not directly comparable—people of somewhat different ages or smoking habits—removing these differences using a mathematical structure that has elements estimated from the data at hand. Investigators often have great faith in their models, a faith that is expressed in the large tasks they expect their models to successfully perform.

© Springer Nature Switzerland AG 2020

P. R. Rosenbaum, *Design of Observational Studies*, Springer Series in Statistics,
https://doi.org/10.1007/978-3-030-46405-9_1

Lawlor et al. then note that a large randomized controlled trial published in the *Lancet* [22] compared a placebo pill with a multivitamin pill including vitamin C, finding slightly but not significantly lower death rates under placebo. The randomized trial and the observational study seem to contradict one another. Why is that? There are, of course, many possibilities. There are some important differences between the randomized trial and the observational study; in particular, the treatments are not really identical, and it is not inconceivable that each study correctly answered questions that differ in subtle ways. In particular, Khaw et al. emphasize vitamin C from fruit and vegetable intake rather than from vitamin supplements. Lawlor et al. examine a different possibility, namely that, because of the absence of randomized treatment assignment, people who were not really comparable were compared in the observational study. Their examination of this possibility is indirect, using data from another study, the British Women's Heart and Health Study, in which several variables were measured that were not included in the adjustments performed by Khaw et al. Lawlor et al. find that women with low levels of vitamin C in their blood are more likely to smoke cigarettes, to exercise less than 1 h per week, to be obese, and less likely to consume a low fat diet, a high fiber diet, and daily alcohol. Moreover, women with low levels of vitamin C in their blood are more likely to have had a childhood in a "manual social class," with no bathroom or hot water in the house, a shared bedroom, no car access, and to have completed full-time education by 18 years of age. And the list goes on. The concern is that one or more of these differences, or some other difference that was not measured, not the difference in vitamin C, is responsible for the higher coronary mortality among individuals with lower levels of vitamin C in their blood. To a large degree, this problem was avoided in the randomized trial, because there, only the turn of a coin distinguished placebo and multivitamin.

1.2 Cochran's Basic Advice

> The planner of an observational study should always ask himself the question, "How would the study be conducted if it were possible to do it by controlled experimentation?"
>
> William G. Cochran [11, page 236]
> attributing the point to H.F. Dorn.

At the most elementary level, a well designed observational study resembles, as closely as possible, a simple randomized experiment. By definition, the resemblance is incomplete: randomization is not used to assign treatments in an observational study. Nonetheless, elementary mistakes are often introduced and opportunities missed by unnecessary deviations from the experimental template. The current section briefly mentions these most basic ingredients.

1.2.1 Treatments, Covariates, Outcomes

Randomized experiment: There is a well-defined treatment, that began at a well-defined time, so there is a clear distinction between covariates measured prior to treatment, and outcomes measured after treatment.

Better observational study: There is a well-defined treatment, that began at a well-defined time, so there is a clear distinction between covariates measured prior to treatment, and outcomes measured after treatment.

Poorer observational study: It is difficult to say when the treatment began, and some variables labeled as covariates may have been measured after the start of treatment, so they might have been affected by the treatment. The distinction between covariates and outcomes is not clear. See [38].

1.2.2 How were Treatments Assigned?

Randomized experiment: Treatment assignment is determined by a truly random device. At one time, this actually meant coins or dice, but today it typically means random numbers generated by a computer.

Better observational study: Treatment assignment is not random, but circumstances for the study were chosen so that treatment seems haphazard, or at least not obviously related to the outcomes subjects would exhibit under treatment or under control. When investigators are especially proud, having found unusual circumstances in which treatment assignment, though not random, seems unusually haphazard, they may speak of a "natural experiment."

Poorer observational study: Little attention is given to the process that made some people into treated subjects and others into controls.

1.2.3 Were Treated and Control Groups Comparable?

Randomized experiment: Although a direct assessment of comparability is possible only for covariates that were measured, a randomized trial typically has a table demonstrating that the randomization was reasonably effective in balancing these observed covariates. Randomization provides some basis for anticipating that many covariates that were not measured will tend to be similarly balanced.

Better observational study: Although a direct assessment of comparability is possible only for covariates that were measured, a matched observational study typically has a table demonstrating that the matching was reasonably effective in balancing these observed covariates. Unlike randomization, matching for observed covariates provides absolutely no basis for anticipating that unmeasured covariates are similarly balanced.

Poorer observational study: No direct assessment of comparability is presented.

1.2.4 Eliminating Plausible Alternatives to Treatment Effects

Randomized experiment: The most plausible alternatives or counterclaims to an actual treatment effect are identified, and the experimental design includes features to shed light on these alternatives. Typical examples include the use of placebos and other forms of sham or partial treatment, or the blinding of subjects and investigators to the identity of the treatment received by a subject.

Better observational study: The most plausible alternatives or counterclaims to an actual treatment effect are identified, and the design of the observational study includes features to shed light on these alternatives. Because there are many more plausible alternatives to a treatment effect in an observational study than in an experiment, much more effort is devoted to collecting data that would shed light on these alternatives. Typical examples include multiple control groups thought to be affected by different biases, or a sequence of longitudinal baseline pretreatment measurements of the variable that will be the outcome after treatment. When investigators are especially proud of devices included to distinguish treatment effects from plausible alternatives, they may speak of a "quasi-experiment."

Poorer observational study: Plausible alternatives to a treatment effect are mentioned in the discussion section of the published report.

1.2.5 Exclusion Criteria

Randomized experiment: Subjects are included or excluded from the experiment based on covariates, that is, on variables measured prior to treatment assignment and hence unaffected by treatment. Only after the subject is included is the subject randomly assigned to a treatment group and treated. This ensures that the same exclusion criteria are used in treated and control groups.

Better observational study: Subjects are included or excluded from the experiment based on covariates, that is, on variables measured prior to treatment assignment and hence unaffected by treatment. The same criteria are used in treated and control groups.

Poorer observational study: A person included in the control group might have been excluded if assigned to treatment instead. The criteria for membership in the treated and control groups differ. In one particularly egregious case, to be discussed in Sect. 13.1, treatment was not immediately available, and any patient who died before the treatment became available was placed in the control group; then came the exciting news that treated patients lived longer than controls.

1.2.6 Exiting a Treatment Group After Treatment Assignment

Randomized experiment: Once assigned to a treatment group, subjects do not exit. A subject who does not comply with the assigned treatment, or switches to another treatment, or is lost to follow-up, remains in the assigned treatment group with these characteristics noted. An analysis that compares the groups as randomly assigned, ignoring deviations between intended and actual treatment, is called an "intention-to-treat" analysis, and it is one of the central analyses reported in a randomized trial. Randomization inference may partially address noncompliance with assigned treatment by viewing treatment assignment as an instrumental variable for treatment received; see Sect. 5.3 and [20].

Better observational study: Once assigned to a treatment group, subjects do not exit. A subject who does not comply with the assigned treatment, or switches to another treatment, or is lost to follow-up, remains in the assigned treatment group with these characteristics noted. Inference may partially address noncompliance by viewing treatment assignment as an instrumental variable for treatment received; see Sect. 5.3 and [24].

Poorer observational study: There is no clear distinction between assignment to treatment, acceptance of treatment, receipt of treatment, or switching treatments, so problems that arise in experiments seem to be avoided, when in fact they are simply ignored.

1.2.7 Study Protocol

Randomized experiment: Before beginning the actual experiment, a written protocol describes the design, exclusion criteria, primary and secondary outcomes, and proposed analyses.

Better observational study: Before examining outcomes that will form the basis for the study's conclusions, a written protocol describes the design, exclusion criteria, primary and secondary outcomes, and proposed analyses; see Chap. 23.

Poorer observational study: If sufficiently many analyses are performed, something publishable will turn up sooner or later.

1.3 Maimonides' Rule

In 1999, Joshua Angrist and Victor Lavy [3] published an unusual and much admired study of the effects of class size on academic achievement. They wrote [3, pages 533–535]:

> [C]ausal effects of class size on pupil achievement have proved very difficult to measure. Even though the level of educational inputs differs substantially both between and within

> schools, these differences are often associated with factors such as remedial training or students' socioeconomic background ... The great twelfth century Rabbinic scholar, Maimonides, interprets the Talmud's discussion of class size as follows: "Twenty-five children may be put in charge of one teacher. If the number in the class exceeds twenty-five but is not more than forty, he should have an assistant to help with instruction. If there are more than forty, two teachers must be appointed." ... The importance of Maimonides' rule for our purposes is that, since 1969, it has been used to determine the division of enrollment cohorts into classes in Israeli public schools.

In most places at most times, class size has been determined by the affluence or poverty of a community, its enthusiasm or skepticism about the value of education, the special needs of students for remedial or advanced instruction, the obscure, transitory, barely intelligible obsessions of bureaucracies, and each of these determinants of class size clouds its actual effect on academic performance. However, if adherence to Maimonides' rule were perfectly rigid, then what would separate a school with a single class of size 40 from the same school with two classes whose average size is 20.5 is the enrollment of a single student.

Maimonides' rule has the largest impact on a school with about 40 students in a grade cohort. With cohorts of size 40, 80, and 120 students, the steps down in average class size required by Maimonides' rule when an additional student enrolls are, respectively, from 40 to 20.5, from 40 to 27, and from 40 to 30.25. For this reason, we will look at schools with fifth grade cohorts in 1991 with between 31 and 50 students, where average class sizes might be cut in half by Maimonides' rule. There were 211 such schools, with 86 of these schools having between 31 and 40 students in fifth grade, and 125 schools having between 41 and 50 students in the fifth grade.

Adherence to Maimonides' rule is not perfectly rigid. In particular, Angrist and Lavy [3, page 538] note that the percentage of disadvantaged students in a school "is used by the Ministry of Education to allocate supplementary hours of instruction and other school resources." Among the 211 schools with between 31 and 50 students in fifth grade, the percentage disadvantaged has a slightly negative Kendall's correlation of -0.10 with average class size, which differs significantly from zero (P-value $= 0.031$), and it has more strongly negative correlations of -0.42 and -0.55, respectively, with performance on verbal and mathematics test scores. For this reason, 86 matched pairs of two schools were formed, matching to minimize the total absolute difference in percentage disadvantaged. Figure 1.1 shows the paired schools, 86 schools with 31 and 40 students in fifth grade, and 86 schools with between 41 and 50 students in the fifth grade. After matching, the upper left panel in Fig. 1.1 shows that the percentage of disadvantaged students was balanced; indeed, the average absolute difference within a pair was less than 1%. The upper right panel in Fig. 1.1 shows Maimonides' rule at work: with some exceptions, the slightly larger schools had substantially smaller class sizes. The bottom panels of Fig. 1.1 show the average mathematics and verbal test performance of these fifth graders, with somewhat higher scores in the schools with between 41 and 50 fifth graders, where class sizes tended to be smaller.

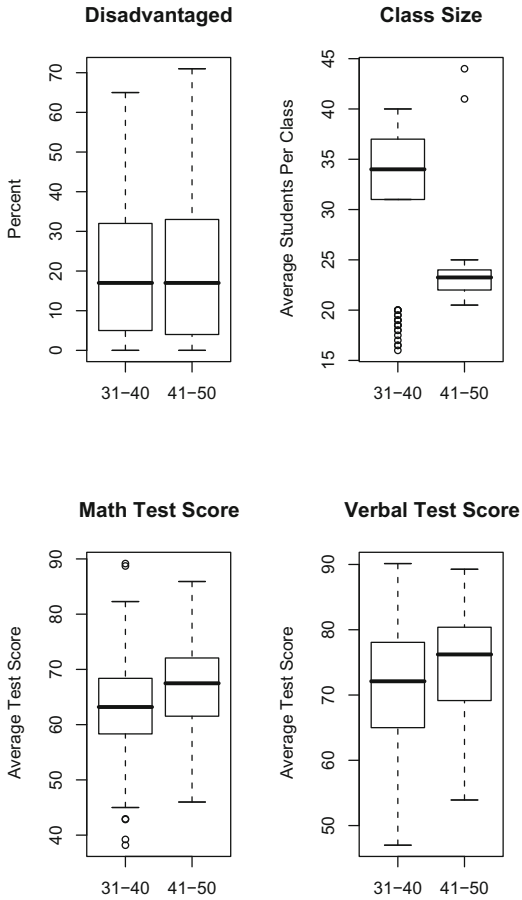

Fig. 1.1 Eighty-six pairs of two Israeli schools, one with between 31 and 40 students in the fifth grade, the other with between 41 and 50 students in the fifth grade, matched for percentage of students in the school classified as disadvantaged. The figure shows that the percentage of disadvantaged students is balanced, that imperfect adherence to Maimonides' rule has yielded substantially different average class sizes, and test scores were higher in the group of schools with predominantly smaller class sizes

1.4 Seat Belts in Car Crashes

Do safety belts prevent fatalities in automobile accidents? Car crashes vary in severity, depending upon speed, road traction, the reaction time of a driver stepping on the brakes, and physical forces that are rarely, if ever, measured. Wearing safety belts is a precaution. Many people, perhaps most people, who wear safety belts think that a serious accident is possible; this possibility is salient, and small inconveniences seem tolerable if the risk is reduced. In contrast, small inconveniences may seem intolerable if a serious accident is seen as a remote possibility. Does one take a single precaution? Perhaps some people do, but others will take several precautions. If cautious drivers wear seat belts, but also drive at slower speeds, at a greater distance from the car ahead, with greater allowance for road conditions—if risk-tolerant drivers do not wear seat belts, drive faster and closer, ignore road conditions—then

Table 1.1 Crashes in FARS 1975–1983 in which the front seat had two occupants, a driver and a passenger, with one belted, the other unbelted, and one died and one survived

	Driver Passenger	Not belted Belted	Belted Not belted
Driver died	Passenger survived	189	153
Driver survived	Passenger died	111	363

a simple comparison of belted and unbelted drivers may credit seat belts with effects that reflect, in part, the severity of the crash.

Using data from the U.S. Fatal Accident Reporting System (FARS), Leonard Evans [16] looked at crashes in which there were two individuals in the front seat, one belted, the other unbelted, with at least one fatality. In these crashes, several, otherwise, uncontrolled features are the same for driver and passenger: speed, road traction, distance from the car ahead, reaction time. Admittedly, risk in the passenger seat may differ from risk in the driver seat, but in this comparison there are belted drivers with unbelted passengers and unbelted drivers with belted passengers, so this issue may be examined. Table 1.1 is derived from Evans' [16] more detailed tables. In this table, when the passenger is belted and the driver is not, more often than not, the driver dies; conversely, when the driver is belted and the passenger is not, more often than not, the passenger dies.

Everyone in Table 1.1 is at least 16 years of age. Nonetheless, the roles of driver and passenger are connected to law and custom, for parents and children, husbands and wives, and others. For this reason, Evans did further analyses, for instance, taking account of the ages of driver and passenger, with similar results.

Evans [16, page 239]wrote:

> The crucial information for this study is provided by cars in which the safety belt use of the subject and other occupant differ ... There is a strong tendency for safety belt use or non-use to be the same for different occupants of the same vehicle ... Hence, sample sizes in the really important cells are ... small ...

This study is discussed further in Sect. 5.2.6.

1.5 Money for College

To what extent, if any, does financial aid increase college attendance? It would not do to simply compare those who received aid with those who did not. Decisions about the allocation of financial aid are often made person by person, with consideration of financial need and academic promise, together with many other factors. A grant of financial aid is often a response to an application for aid, and the decision to apply or not is likely to reflect an individual's motivation for continued education and competing immediate career prospects.

To estimate the effect of financial aid on college attendance, Susan Dynarski [15] used "a shift in aid policy that affect[ed] some students but not others." Between 1965 and 1982, a program of the U.S. Social Security Administration provided substantial financial aid to attend college for the children of deceased Social Security beneficiaries, but the U.S. Congress voted in 1981 to end the program. Using data from the National Longitudinal Survey of Youth, Dynarski [15] compared college attendance of high school seniors with deceased fathers and high school seniors whose fathers were not deceased, in 1979–1981 when aid was available, and in 1982–1983 after the elimination of the program. Figure 1.2 depicts the comparison. In 1979–1981, while the Social Security Student Benefit Program provided aid to students with deceased fathers, these students were more likely than others to attend college, but in 1982–1983, after the program was eliminated, these students were less likely than others to attend college.

In Fig. 1.2, the group that faced a change in incentives exhibited a change in behavior, whereas the group that faced no change in incentives exhibited little change in behavior. In the spirit of Sect. 1.2.4, Fig. 1.2 studies one treatment using four groups, where only certain patterns of response among the four groups are compatible with a treatment effect; see also [9, 41] and [58, Chapter 5].

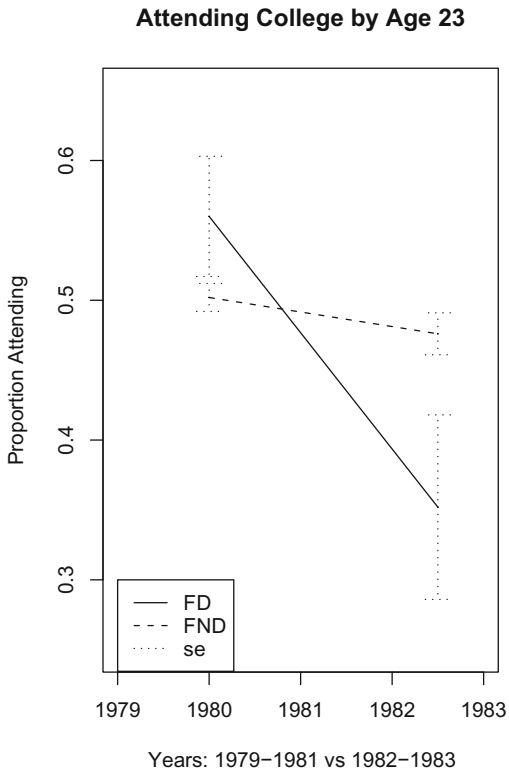

Fig. 1.2 College attendance by age 23 in four groups: before (1979–1981) and after (1982–1983) the end of the Social Security Student Benefit Program for children whose fathers were deceased (FD) or not deceased (FND). Values are proportions with standard errors (se)

Is being the child of a deceased father a random event? Apparently not. It is unrelated to the child's age and gender, but the children of deceased fathers had mothers and fathers with less education and were more likely to be black; however, these differences were about the same in 1979–1981 and 1982–1983, so these differences alone are not good explanations of the shift in college attendance [15, Table 1]. This study is discussed further in Chap. 14.

1.6 Nature's "Natural Experiment"

In asking whether a particular gene plays a role in causing a particular disease, a key problem is that the frequencies of various forms of a gene (its alleles) vary somewhat from one human community to the next. At the same time, habits, customs, diets, and environments also vary somewhat from one community to the next. In consequence, an association between a particular allele and a particular disease may not be causal: gene and disease may both be associated with some cause, such as diet, that is not genetic. Conveniently, nature has created a natural experiment.

With the exception of sex-linked genes, a person receives two versions of each gene, perhaps identical, one from each parent, and transmits one copy to each child. To a close approximation, in the formation of a fertilized egg, each parent contributes one of two possible alleles, each with probability $\frac{1}{2}$, the contributions of the two parents being independent of each other, and independent for different children of the same parents. (The transmissions of different genes that are neighbors on the same chromosome are not generally independent; see [61, §15.4]. In consequence, a particular gene may be associated with a disease not because it is a cause of the disease, but rather because it is a marker for a neighboring gene that is a cause.)

Several strategies use this observation to create natural experiments that study genetic causes of a specific disease. Individuals with the disease are identified. Richard Spielman, Ralph McGinnis, and Warren Ewens [60] used genetic information on the diseased individual and both parents in their Transmission/Disequilibrium Test (TDT). The test compares the diseased individuals to the known distributions of alleles for the hypothetical children their parents could produce. David Curtis [14], Richard Spielman and Warren Ewens [59], and Michael Boehnke and Carl Langefeld [7] suggested using genetic information on the diseased individual and one or more siblings from the same parents, which Spielman and Ewens called the sib-TDT. If the disease has no genetic cause linked to the gene under study, then the alleles from the diseased individual and her siblings should be exchangeable.

The idea underlying the sib-TDT is illustrated in Table 1.2, using data from Boehnke and Langefeld [7, Table 5], their table being derived from work of Margaret Pericak-Vance and Ann Saunders; see [53]. Table 1.2 gives the frequency of the ϵ_4 allele of the apolipoprotein E gene in 112 individuals with Alzheimer

Table 1.2 Alzheimer disease and the apolipoprotein E ϵ_4 allele in 112 sibling pairs, one with Alzheimer disease (affected), the other without (unaffected)

		Unaffected Sib		
	# ApoEϵ_4 Alleles	0	1	2
Affected Sib	0	23	4	0
	1	25	36	2
	2	8	8	6

The table counts pairs, not individuals. The rows and columns of the table indicate the number (0, 1, or 2) of ApoE alleles for the affected and unaffected sibling

disease and in an unaffected sibling of the same parents. Table 1.2 counts sibling pairs, not individuals, so the total count in the table is 112 pairs of an affected and an unaffected sibling. Each person can receive 0, 1, or 2 copies of the ϵ_4 allele from parents. For any one pair, write (aff, unaff) for the number of ϵ_4 alleles possessed by, respectively, the affected and unaffected sibling. In Table 1.2, there are 25 pairs with $(\text{aff}, \text{unaff}) = (1, 0)$. If Alzheimer disease had no genetic link with the apolipoprotein E ϵ_4 allele, then nature's natural experiment implies that the chance that $(\text{aff}, \text{unaff}) = (2, 0)$, say, is the same as the chance that $(\text{aff}, \text{unaff}) = (0, 2)$, and more generally, the chance that $(\text{aff}, \text{unaff}) = (i, j)$ equals the chance that $(\text{aff}, \text{unaff}) = (j, i)$, for $i, j = 0, 1, 2$. In fact, this does not appear to be the case in Table 1.2. For instance, there are eight sibling pairs such that $(\text{aff}, \text{unaff}) = (2, 0)$ and none such that $(\text{aff}, \text{unaff}) = (0, 2)$. Also, there are 25 pairs such that $(\text{aff}, \text{unaff}) = (1, 0)$ and only 4 pairs such that $(\text{aff}, \text{unaff}) = (0, 1)$.

A distribution with the property

$$\Pr\{(\text{aff}, \text{unaff}) = (i, j)\} = \Pr\{(\text{aff}, \text{unaff}) = (j, i)\} \quad \text{for all } i, \ j$$

is said to be exchangeable. In the absence of a genetic link with disease, nature's natural experiment ensures that the distribution of allele frequencies in affected/unaffected sib pairs is exchangeable. This creates a test, the sib transmission disequilibrium test [59] that is identical to a certain randomization test appropriate in a randomized experiment [33].

1.7 What This Book Is About

1.7.1 Basic Structure

Design of Observational Studies has five parts, "Beginnings," "Matching," "Design Sensitivity," "Enhanced Design," and "Planning Analysis," plus a brief summary. Part I, "Beginnings," is a conceptual introduction to causal inference in observational studies. Chapters 2, 3, and 5 of Part I cover concisely, in about one hundred pages, many of the ideas discussed in my book *Observational Studies* [42], but in a far less technical and less general fashion. Parts II–V cover material that,

for the most part, has not previously appeared in book form. Part II, "Matching," concerns the conceptual, practical, and computational aspects of creating a matched comparison that balances many observed covariates. Because matching does not make use of outcome information, it is part of the design of the study, what Cochran [11] called "setting up the comparisons"; that is, setting up the structure of the experimental analog. Even if the matching in Part II is entirely successful, so that after matching, matched treated and control groups are comparable with respect to all observed covariates, the question or objection or challenge will inevitably be raised that subjects who look comparable in observed data may not actually be comparable in terms of covariates that were not measured. Chapters 3 and 5 and Parts III, IV and V address this central concern. Part III, "Design Sensitivity," discusses a quantitative tool for appraising how well competing designs (or data generating processes) resist such challenges. In part, "Design Sensitivity" will provide a formal appraisal of the design strategies introduced informally in Chap. 5. Part IV discusses further aspects of design using concepts from Part III. Part V discusses those activities that follow matching but precede analysis, notably planning the analysis.

1.7.2 Structure of Part I: Beginnings

Observational studies are built to resemble simple experiments, and Chap. 2 reviews the role that randomization plays in experiments. Chapter 2 also introduces elements and notation shared by experiments and observational studies. Chapter 3 discusses two simple models for observational studies, one claiming that adjustments for observed covariates suffice, the other engaging the possibility that they do not. Chapter 3 introduces the propensity score and sensitivity analysis. Observational studies are built from three basic ingredients: opportunities, devices, and instruments. Chapter 5 introduces these ideas in an informal manner, with some of the formalities developed in Part III and others developed in [42, Chapters 4, 6–9].

It is typical, perhaps inevitable, that an observational study is greeted with objections and counterclaims. This is the topic of Chap. 7.

My impression is that many observational studies dissipate either by the absence of a focused objective or by becoming mired in ornate analyses that may overwhelm an audience but are unlikely to convince anyone. Neither problem is common in randomized experiments, and both problems are avoidable in observational studies. Chapter 4 discusses the first problem, while Chap. 6 discusses the second. In a successful experiment or observational study, competing theories make conflicting predictions; this is the concern of Chap. 4. Transparency means making evidence evident, and Chap. 6 discusses how this is done.

1.7.3 Structure of Part II: Matching

Part II, entitled "Matching," is partly conceptual, partly algorithmic, partly data analytic. Chapter 8 is introductory: it presents a matched comparison as it might (and did) appear in a scientific journal. The basic tools of multivariate matching are described and illustrated in Chap. 9, and various common practicalities are discussed in Chap. 10. Later chapters in Part II discuss specific topics in matching, including fine balance, matching with multiple groups or without groups, and risk-set matching. Matching in the computer package `R` is discussed in Chap. 14.

1.7.4 Structure of Part III: Design Sensitivity

In Chap. 3, it is seen that some observational studies are sensitive to small unobserved biases, whereas other studies are insensitive to quite large unobserved biases. What features of the design of an observational study affect its sensitivity to bias from covariates that were not measured? This is the focus of Part III.

Chapter 15 reviews the concept of power in a randomized experiment, then defines the power of a sensitivity analysis. Design sensitivity is then defined. Design sensitivity is a number that defines the sensitivity of an observational study design to unmeasured biases when the sample size is large. Many factors affect the design sensitivity, including the issues discussed informally in Chap. 5. Chapter 16 revives a very old debate between John Stuart Mill and Sir Ronald Fisher about the relevance to causal inference of the heterogeneity of experimental material. Mill believed it mattered quite a bit; Fisher denied this. Sometimes a treatment has little effect on most people and a dramatic effect on some people. In one sense, the effect is small—on average it is small—but for a few people it is large. Is an effect of this sort highly sensitive to unmeasured biases? Chapter 17 provides the answer. Chapter 18 takes up themes from Chap. 5, specifically coherence and dose–response, and evaluates their contribution to design sensitivity.

Although the study design affects design sensitivity, so do the methods of analysis. This large topic is sketched in Chap. 19.

1.7.5 Structure of Part IV: Enhanced Design

Part IV discusses two topics in study design that make use of ideas from Part III. These chapters are linked to Chap. 5, but they are more technical.

A study has two evidence factors if it permits two statistically independent tests of the hypothesis of no treatment effect, where biases that would invalidate the first test have no effect on the second. Evidence factors are discussed in Chap. 20.

In Chaps. 5, 12, and 20, comparisons among several groups—not just a treated and a control group—provide strengthened evidence that a treatment actually causes its ostensible effects. Chapter 21 discusses the construction of these groups.

1.7.6 Structure of Part V: Planning Analysis

The sample has been successfully matched—treated and control groups look comparable in terms of measured covariates—and Part IV turns to planning the analysis. Chapter 22 concerns three empirical steps that aid in planning the analysis: sample splitting to improve design sensitivity, checking that analytical adjustments are feasible, and thick description of a few matched pairs. After reviewing Fisher's advice—"make your theories elaborate"–Chap. 23 discusses planning the analysis of an observational study.

1.7.7 A Less Technical Introduction to Observational Studies

The mathematician Paul Halmos wrote two essays, "How to write mathematics" and "How to talk mathematics." In the latter, he suggested that in a good mathematical talk, you do not prove everything, but you do prove something to give the flavor of proofs for the topic under discussion. In the spirit of that remark, *Observational Studies* [42] writes about statistics, where *Design of Observational Studies* talks about statistics. This is done in several ways.

We often develop an understanding by taking something apart and putting it back together. In statistics, this typically means looking at an unrealistically small example in which it is possible to see the details of what goes on. For this reason, I discuss several unrealistically small examples in parallel with real examples of practical size. For instance, Chap. 2 discusses two versions of a paired randomized experiment, one with five pairs, the other with 185 pairs. The five pairs are a random sample from the 185 pairs. With 5 pairs, there are $2^5 = 32$ possible treatment assignments, and it is possible to see what is happening. With 185 pairs, there are $2^{185} = 4.9 \times 10^{55}$ possible treatment assignments, and it is not possible to see what is happening, although nothing new is happening beyond what you saw happening with five pairs. The larger experiment is just larger, somewhat awkward to inspect, but no different conceptually. In a similar way, Chap. 8 discusses the construction of 344 pairs matched for many covariates, while Chap. 9 discusses the construction of 21 pairs matched for three covariates. With 21 pairs, you can see what is happening, whereas with 344 pairs you cannot see as much, but nothing new is happening.

Chapter 2 discusses a number of very old, very central concepts in statistics. These include: the role of randomization in experiments, the nature of randomization tests, obtaining confidence intervals by inverting hypothesis tests, building an estimator using an estimating equation, and so on. This material is so old and central to the subject that an article in a statistical journal might reduce the entire chapter to a paragraph, and that would be fine for someone who had been around this track a few times. My goal in Chap. 2 is not concise expression. My goal in Chap. 2 is to take you around this track a few times.

To prove something, not everything, I develop statistical theory only for the case of matched pairs with continuous responses. The case of matched pairs is the simplest nontrivial case. All of the important concepts appear in the case of matched pairs, but most of the technical issues are easy. Randomization distributions for matched pairs are little more than a series of independent coin flips. Everybody can do coin flips. In a randomized experiment, the coin flips are fair, but in a sensitivity analysis, the coin flips may be biased. The matching methods in Part II are not restricted to pair matching—matching with multiple controls, matching with variable controls, full matching, risk-set matching are all there—however, if you want to work through the derivations of the associated statistical analyses for continuous, discrete, censored, and multivariate responses, you will need to turn to [42] or the references discussed in "Further Reading."

Focusing the theoretical presentation on matched pairs permits discussion of key concepts with the minimum of mathematics. Unlike statistical analysis, research design yields decisions rather than calculations—decisions to ask certain questions, in certain settings, collecting certain data, adding certain design elements, attending to certain patterns—and for such decisions, the concepts are more important than details of general computations. What is being left out by focusing on matched pairs? For one thing, sensitivity analyses in other cases are easy to do but require more mathematical machinery to justify. Some of this machinery is aesthetically pleasing, for instance, exact results using Holley's inequality [1, 8, 27] in I. R. Savage's [54] finite distributive lattice of rank orders or samples; see [39, 40] and [42, §4]. Some of this machinery uses large sample approximations or asymptotics that work easily and well even in small samples, but discussion of these approximations means a step up in the level of technical detail; see [18, 48] and [42, §4]. To get a feeling for the difference between the paired case and other cases, see [44, 48], where both cases are discussed in parallel. If you need to do a sensitivity analysis for a situation other than matched pairs, see [42, §4] or other references in Further Reading. Another item that is left out of the current discussion is a formal notation and model for multiple groups, as opposed to a treated and a control group; see [42, §8] for such a notation and model. Such a model adds more subscripts and symbols with few additional concepts. The absence of such a notation and model has a small effect on the discussion of multiple control groups (Sect. 5.2.2), differential effects (Sect. 5.2.6), the presentation of the matching session in `R` for a difference-in-differences example in Chap. 14, and the planned analysis with two control groups (Sect. 23.3); specifically, these topics are described informally with reference to the literature for formal results.

1.7.8 Dependence Among Chapters

Design of Observational Studies is highly modular, so it is not necessary to read chapters in order. Part II may be read before or after Part I or not at all; Part II is not

needed for Parts III and V. Part III may be read before or after Part V. Part IV may be read before or after Part V.

In Part I, Chap. 5 depends on Chap. 3, which in turn depends on Chap. 2. The beginning of Chap. 2, up through Sect. 2.4.3, is needed for Chaps. 3 and 5, but Sect. 2.5 is not essential except for Chap. 17, and the remainder of Chap. 2 is not used later in the book. Chapters 4 and 6 may be read at any time or not at all.

In Part II, most chapters depend strongly on Chap. 9 but only weakly on each other. Read the introductory Chaps. 8 and 9; then, read what you like in Part II.

The situation is similar in Part III. All of the chapters of Part III depend upon Chap. 15, which in turn depends on Chap. 5. The remaining chapters of Part III may be read in any order or not at all.

The two chapters of Part IV may be read out of sequence. Chapter 20 depends upon Chap. 19. The appendix to Chap. 20 is considerably more technical than the chapter itself. Chapter 21 depends strongly on Part II, weakly on Part III.

The two chapters in Part V may be read out of sequence, and both depend upon Chap. 5.

At the back of the book, there is a list of symbols and a glossary of statistical terms. In the index, a **bold page number** locates the definition of a technical term or symbol.

Some books (e.g., [42]) contain practice problems for you to solve. My sense is that the investigator planning an observational study has problems enough, so instead of further problems, at the back of the book there is a list of solutions.

As a scholar at a research university, I fall victim to periodic compulsions to make remarks that are largely unintelligible and totally unnecessary. These remarks are found in appendices and footnotes. Under no circumstances read them. If you read a footnote, you will suffer a fate worse than Lot's wife.[1]

1.8 Further Reading

One might reasonably say that the distinction between randomized experiments and observational studies was introduced by Sir Ronald Fisher's [17] invention of randomized experimentation. Fisher's book [17] of 1935 is of continuing interest. As noted in Sect. 1.2, William Cochran [11] argued that observational studies should be understood in relation to experiments; see also the important paper in this spirit by Donald Rubin [50]. A modern discussion of quasi-experiments is given by William Shadish, Thomas Cook, and Donald Campbell [58], and Campbell's [10] collected papers are of continuing interest. See also [19, 35, 64]. Natural experiments in medicine are discussed by Jan Vandenbroucke [63] and the report

[1]She turned into a pillar of salt when she looked where her husband instructed her not to look. Opposed to the story of Lot and his wife is the remark of Immanuel Kant: "Sapere aude" (Dare to know) [29, page 17]. You are, by the way, off to a bad start with these footnotes.

edited by Michael Rutter [52], and in economics by Joshua Angrist and Alan Kruger [2], Timothy Besley and Anne Case [6], Daniel Hamermesh [21], Bruce Meyer [35], and Mark Rosenzweig and Kenneth Wolpin [49]. Natural experiments are prominent also in recent developments in genetic epidemiology [7, 14, 33, 59, 60], and in other fields [4, 56]. The papers by Jerry Cornfield and colleagues [13], Austin Bradford Hill [25], and Mervyn Susser [62] remain highly influential in epidemiology and are of continuing interest. Miguel Hernán and colleagues [24] illustrate the practical importance of adhering to the experimental template in designing an observational study. For a general discussion of observational studies, see [42, 47].

Why focus on matched pairs in this book? Why does the case of matched pairs lead to simpler mathematics? How are pairs different from matched sets or strata? The following slightly technical explanation may be skipped. In the case of matched pairs, the sensitivity bound in Chap. 3 is provided by a statistic that is a bound *as a random variable*: it is largest, or smallest, as a random variable. Here, the largest random variable means largest in the sense of stochastic order: random variable A is larger than or equal to B in the sense of stochastic order if $\Pr(A \geq c) \geq \Pr(B \geq c)$ at *every* cutpoint c. Unlike the ordering of the real numbers, stochastic ordering of random variables is merely a partial ordering: it may happen that random variable A is not larger than or equal to B, yet B is also not larger than or equal to A; so, neither is larger, nor are they equal. For instance, it could happen that $\Pr(A \geq 0) > \Pr(B \geq 0)$, but $\Pr(A \geq 5) < \Pr(B \geq 5)$, perhaps because A is tightly concentrated near 1, while B has median -1 but is widely dispersed on the interval $[-100, \ 100]$. The set of Normal distributions with variance 1 and mean μ, $-1 \leq \mu \leq 1$, is ordered by μ, and the Normal distribution with $\mu = 1$ is the largest distribution in this set in the sense of stochastic order. In contrast, the Normal distribution with mean $\mu = 0$ and variance $\sigma^2 = 1$ is neither larger nor smaller than the Normal distribution with $\mu = 1$ variance $\sigma^2 = 100$ in the sense of stochastic order. The case of matched pairs is simpler than the general case because we can pick out one random variable that determines the upper sensitivity bound on inference quantities, such as P-values, estimates, and the endpoints of a confidence intervals. This same simplification also occurs without matched pairs when the outcome is binary, or when the test statistic cuts a continuous outcome to make it binary; see [42, §4.4]. In the general case, however, one obtains sensitivity bounds by solving an optimization problem defined by the data at hand, for instance, maximizing a P-value or an estimate. The general case is slightly less tidy as a mathematical problem [44, 48], but it is easy to handle in practice with appropriate software, as discussed in the next paragraph. By and large, the concepts involved in the design of observational studies may be discussed in the case of matched pairs, thereby avoiding this slightly less tidy mathematics.[2]

What should I do if I do not have matched pairs? First, understand the paired case; it is easy to understand. Though slightly less tidy [44, 48], the general case is easy to handle in data analysis using appropriate software [46], such as

[2] An exception is the observation that, with continuous outcomes, the presence of multiple matched controls affects the design sensitivity; see [45] and [47, pp. 222–223].

the `senm` and `senmCI` functions in the `sensitivitymult` package in `R` for multiple controls [44], the `sensitivityfull` package for full matching, or the `senstrat` package for stratified comparisons [48]. Moreover, the general solution may combine stratification for covariates with model-based covariance adjustment for covariates; see [43, 48]. For general discussion, see [42, §4].

Bibliography

1. Anderson, I.: Combinatorics of Finite Sets. Oxford University Press, New York (1987)
2. Angrist, J.D., Krueger, A.B.: Empirical strategies in labor economics. In: Ashenfelter, O., Card, D. (eds.) Handbook of Labor Economics, vol. 3, pp. 1277–1366. Elsevier, New York (1999)
3. Angrist, J.D., Lavy, V.: Using Maimonides' rule to estimate the effect of class size on scholastic achievement. Q. J. Econ. **114**, 533–575 (1999)
4. Armstrong, C.S., Kepler, J.D.: Theory, research design assumptions, and causal inferences. J. Account. Econ. **66**, 366–373 (2018)
5. Athey, S., Imbens, G. W.: The state of applied econometrics: causality and policy evaluation. J. Econ. Perspect. **31**, 3–32 (2018)
6. Besley, T., Case, A.: Unnatural experiments? Estimating the incidence of endogenous policies. Econ. J. **110**, 672–694 (2000)
7. Boehnke, M., Langefeld, C.D.: Genetic association mapping based on discordant sib pairs: the discordant alleles test. Am. J. Hum. Genet. **62**, 950–961 (1998)
8. Bollobás, B.: Combinatorics. Cambridge University Press, New York (1986)
9. Campbell, D.T.: Factors relevant to the validity of experiments in social settings. Psychol. Bull. **54**, 297–312 (1957)
10. Campbell, D.T.: Methodology and Epistemology for Social Science: Selected Papers. University of Chicago Press, Chicago (1988)
11. Cochran, W.G.: The planning of observational studies of human populations (with Discussion). J. R. Stat. Soc. Ser. A **128**, 234–265 (1965)
12. Cook, T.D., Shadish, W.R.: Social experiments: some developments over the past fifteen years. Annu. Rev. Psychol. **45**, 545–580 (1994)
13. Cornfield, J., Haenszel, W., Hammond, E., Lilienfeld, A., Shimkin, M., Wynder, E.: Smoking and lung cancer: recent evidence and a discussion of some questions. J. Natl. Cancer Inst. **22**, 173–203 (1959)
14. Curtis, D.: Use of siblings as controls in case-control association studies. Ann. Hum. Genet. **61**, 319–333 (1997)
15. Dynarski, S.M.: Does aid matter? Measuring the effect of student aid on college attendance and completion. Am. Econ. Rev. **93**, 279–288 (2003)
16. Evans, L.: The effectiveness of safety belts in preventing fatalities. Accid. Anal. Prev. **18**, 229–241 (1986)
17. Fisher, R.A.: Design of Experiments. Oliver and Boyd, Edinburgh (1935)
18. Gastwirth, J.L., Krieger, A.M., Rosenbaum, P.R.: Asymptotic separability in sensitivity analysis. J. R. Stat. Soc. Ser. B **62**, 545–555 (2000)
19. Greenstone, M., Gayer, T.: Quasi-experimental and experimental approaches to environmental economics. J. Environ. Econ. Manag. **57**, 21–44 (2009)
20. Greevy, R., Silber, J.H., Cnaan, A. , Rosenbaum, P.R.: Randomization inference with imperfect compliance in the ACE-inhibitor after anthracycline randomized trial. J. Am. Stat. Assoc. **99**, 7–15 (2004)
21. Hamermesh, D.S.: The craft of labormetrics. Ind. Labor Relat. Rev. **53**, 363–380 (2000)

22. Heart Protection Study Collaborative Group.: MRC/BHF Heart Protection Study of antioxidant vitamin supplementation in 20,536 high-risk individuals: a randomised placebo-controlled trial. Lancet **360**, 23–33 (2002)
23. Heckman, J.J.: Micro data, heterogeneity, and the evaluation of public policy: Nobel lecture. J. Polit. Econ. **109**, 673–748 (2001)
24. Hernán, M.A., Alonso, A., Logan, R., Grodstein, F., Michels, K.B., Willett, W.C., Manson, J.E., Robins, J.M.: Observational studies analyzed like randomized experiments: an application to postmenopausal hormone therapy and coronary heart disease (with Discussion). Epidemiology **19**,766–793 (2008)
25. Hill, A.B.: The environment and disease: association or causation? Proc. R. Soc. Med. **58**, 295–300 (1965)
26. Holland, P.W.: Statistics and causal inference. J. Am. Stat. Assoc. **81**, 945–960 (1986)
27. Holley, R.: Remarks on the FKG inequalities. Commun. Math. Phys. **36**, 227–231 (1974)
28. Imbens, G.W., Wooldridge, J.M.: Recent developments in the econometrics of program evaluation. J. Econ. Lit. **47**, 5–86 (2009)
29. Kant, I.: What is enlightenment? In: Kant, I. (ed.) Toward Perpetual Peace and Other Writings. Yale University Press, New Haven (1785, 2006)
30. Katan, M.B.: Apolipoprotein E isoforms, serum cholesterol, and cancer. Lancet **1**, 507–508 (1986) Reprinted: Int J Epidemiol **33**, 9 (2004)
31. Katan, M.B.: Commentary: mendelian randomization, 18 years on. Int. J. Epidemiol. **33**, 10–11 (2004)
32. Khaw, K.T., Bingham, S., Welch, A., Luben, R., Wareham, N., Oakes, S., Day, N.: Relation between plasma ascorbic acid and mortality in men and women in EPIC-Norfolk prospective study. Lancet **357**, 657–663 (2001)
33. Laird, N.M.,Blacker, D., Wilcox, M.: The sib transmission/disequilibrium test is a Mantel-Haenszel test. Am. J. Hum. Genet. **63**, 1915 (1998)
34. Lawlor, D.A., Smith, G.D., Bruckdorfer, K.R., Kundo, D., Ebrahim, S.: Those confounded vitamins: what can we learn from the differences between observational versus randomized trial evidence? Lancet **363**, 1724–1727 (2004)
35. Meyer, B.D.: Natural and quasi-experiments in economics. J. Bus. Econ. Stat. **13**, 151–161 (1995)
36. Musci, R.J., Stuart, E: Ensuring causal, not casual, inference. Prev. Sci. **20**, 452–456 (2019)
37. Pearce, N., Vandenbroucke, J.P., Lawlor, D.A.: Causal inference in environmental epidemiology: old and new approaches. Epidemiol **30**, 311–316 (2019)
38. Rosenbaum, P.R.: The consequences of adjustment for a concomitant variable that has been affected by the treatment. J. R. Stat. Soc. Ser. A **147**, 656–666 (1984)
39. Rosenbaum, P.R.: On permutation tests for hidden biases in observational studies: an application of Holley's inequality to the Savage lattice. Ann. Stat. **17**, 643–653 (1989)
40. Rosenbaum, P.R.: Quantiles in nonrandom samples and observational studies. J. Am. Stat. Assoc. **90**, 1424–1431 (1995)
41. Rosenbaum, P.R.: Stability in the absence of treatment. J. Am. Stat. Assoc. **96**, 210–219 (2001)
42. Rosenbaum, P.R.: Observational Studies (2nd ed.). Springer, New York (2002)
43. Rosenbaum, P.R.: Covariance adjustment in randomized experiments and observational studies. Stat. Sci. **17**, 286–327 (2002)
44. Rosenbaum, P.R.: Sensitivity analysis for m-estimates, tests, and confidence intervals in matched observational studies. Biometrics **63**, 456–464 (2007)
45. Rosenbaum, P.R.: Impact of multiple matched controls on design sensitivity in observational studies. Biometrics **69**, 118–127 (2013)
46. Rosenbaum, P.R.: Two R packages for sensitivity analysis in observational studies. Obs. Stud. **1**, 1–17 (2015)
47. Rosenbaum, P.R.: Observation and Experiment: An Introduction to Causal Inference. Harvard University Press, Cambridge (2017)
48. Rosenbaum, P.R.: Sensitivity analysis for stratified comparisons in an observational study of the effect of smoking on homocysteine levels. Ann. Appl. Stat. **12**, 2312–2334 (2018)

49. Rosenzweig, M.R., Wolpin, K.I.: Natural 'natural experiments' in economics. J. Econ. Lit. **38**, 827–874 (2000)
50. Rubin, D.B.: Estimating causal effects of treatments in randomized and nonrandomized studies. J. Educ. Psychol. **66**, 688–701 (1974)
51. Rubin, D.B.: The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. Stat. Med. **26**, 20–36 (2007)
52. Rutter, M.: Identifying the Environmental Causes of Disease: How Do We Decide What to Believe and When to Take Action? Academy of Medical Sciences, London (2007)
53. Saunders, A.M., Strittmatter, W.J., Schmechel, D., et al.: Association of apolipoprotein E allele epsilon 4 with late-onset familial and sporadic Alzheimer's disease. Neurology **43**, 1467–1472 (1993)
54. Savage, I.R.: Contributions to the theory of rank order statistics: applications of lattice theory. Revue de l'Institut Int. de Statistique **32**, 52–63 (1964)
55. Sekhon, J.S.: Opiates for the matches: matching methods for causal inference. Annu. Rev. Polit. Sci. **12**, 487–508 (2009)
56. Sekhon, J.S., Titiunik, R.: When natural experiments are neither natural nor experiments. Am. Polit. Sci. Rev. **106**, 35–57 (2012)
57. Shadish, W.R., Cook, T.D.: The renaissance of field experimentation in evaluating interventions. Annu. Rev. Psychol. **60**, 607–629 (2009)
58. Shadish, W.R., Cook, T.D., Campbell, D.T.: Experimental and Quasi-Experimental Designs for Generalized Causal Inference. Houghton-Mifflin, Boston (2002)
59. Spielman, R.S., Ewens, W.J.: A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test. Am. J. Hum. Genet. **62**, 450–458 (1998)
60. Spielman, R.S., McGinnis, R.E., Ewens, W.J.: Transmission test for linkage disequilibrium. Am. J. Hum. Genet. **52**, 506–516 (1993)
61. Strachan, T., Read, A.P.: Human Molecular Genetics. Garland, New York (2004)
62. Susser, M.: Epidemiology, Health and Society: Selected Papers. Oxford University Press, New York (1987)
63. Vandenbroucke, J.P.: When are observational studies as credible as randomized trials? Lancet **363**, 1728–1731 (2004)
64. West, S.G., Duan, N., Pequegnat, W. , Gaist, P. , Des Jarlais, D.C., Holtgrave, D. , Szapocznik, J., Fishbein, M., Rapkin, B., Clatts, M., Mullen, P.D.: Alternatives to the randomized controlled trial. Am. J. Public Health **98**, 1359–1366 (2008)

Chapter 2
Causal Inference in Randomized Experiments

Abstract An observational study is an empiric investigation of treatment effects when random assignment to treatment or control is not feasible. Because observational studies are structured to resemble simple randomized experiments, an understanding of the role randomization plays in experiments is important as background. As a prelude to the discussion of observational studies in later chapters, the current chapter contains a brief review of the logic of causal inference in a randomized experiment. Only one simple case is discussed in detail, namely a randomized paired experiment in which subjects are paired before randomization and one subject in each pair is picked at random to receive treatment, the other receiving control. Although a foundation for later chapters, much of the material in this chapter is quite old, dating from Sir Ronald Fisher's work in the 1920s and 1930s, and it is likely to be familiar from other contexts, such as a course in the design of experiments.

2.1 Two Versions of the National Supported Work Experiment

2.1.1 A Version with 185 Pairs and a Version with 5 Pairs

Though discussed daily in the newspaper, unemployment is a curious phenomenon. Defined abstractly, it implies that someone entered an active labor market intending to sell their labor but was unable to find a buyer, often for long periods of time. Of course, the abstract definition leaves out most of what is happening.

Robert LaLonde [42] reviews several decades of

> public sector sponsored employment and training programs … [intended to] enhance participants' productive skills and, in turn, increase their future earnings and tax payments and reduce their dependence on social welfare benefits. […The] primary recipients of public sector sponsored training [have been] economically disadvantaged or dislocated workers. [42, page 149]

© Springer Nature Switzerland AG 2020

P. R. Rosenbaum, *Design of Observational Studies*, Springer Series in Statistics,
https://doi.org/10.1007/978-3-030-46405-9_2

The National Supported Work Demonstration (NSW) included a randomized experiment evaluating the effects of one such program [11, 36]. Kenneth Couch writes:

> The NSW provided work experiences primarily in service occupations for females and construction for males. The jobs were designed to be consistent with a concept known as graduated stress. Stress within the working environment increased gradually during the training period until it simulated the workplace norms of the private sector. At that point, not more than 18 month after entry, individuals who received services provided by the NSW had to attempt a transition to unsubsidized employment. Screening criteria for the NSW limited participation to individuals severely handicapped in the labor market. ... After screening by a local social service agency and referral into the program, each participant was randomly assigned to a treatment (experimental) or control group. Treatments received the services offered by the NSW. Controls continued enrollment in other available social programs [11, pages 381–382].

The current chapter will use pieces of the NSW experiment to illustrate the logic of randomized experimentation. A reader interested not in randomized experimentation but rather in the full NSW program and its effects might begin with Couch's [11] study of the program's long-term effects. The NSW became important in thinking about methodology for observational studies in economics because of a series of studies [17, 31, 41, 75], beginning with LaLonde's 1986 study [41], in which the randomized control group was set aside and various analytical methods were applied to nonrandomized controls from survey data.

Because the goal is to discuss the logic of randomized experiments in a manner that prepares for the discussion of observational studies in later chapters, several adjustments and simplifications were made. First, the data set is what Rajeev Dehejia and Sadek Wahba [17, Table 1] call the "RE74 subset," which in turn is a subset of the data used by LaLonde [41]. The subset consists of males who were randomized to treatment or control after December 1975 and who had left the program before January 1978, with annual earnings recorded in 1974, 1975, and 1978. Because of these requirements, earnings in 1974 and 1975 are pretreatment covariates unaffected by the NSW program, and earnings in 1978 is an outcome that may be affected by treatment. Furthermore, to emphasize the parallel with matched observational studies, the randomized treated and control groups were matched: specifically, all 185 treated men were matched to 185 untreated controls using eight covariates, forming 185 matched pairs.[1]

[1] As mentioned in Chap. 1, to keep the theoretical technicalities to a minimum, statistical theory is developed only for the case of matched pairs, so I have slightly reshaped the NSW experiment into a matched pair study. Randomization inference for unmatched randomized experiments is similar and is discussed in [59, Chapter 2]. The matching used precisely the methods described later in Sect. 9.4, using a penalty function to implement calipers on the propensity score, with the rank-based Mahalanobis distance used within calipers. Multivariate matching before randomization can improve the efficiency of a randomized experiment; see [29]. For illustration only in this expository chapter, the matching is done after random assignment, which entails discarding some of the randomized controls, something I would not do if the goal were to perform the most efficient analysis of the NSW experiment.

Table 2.1 Pretreatment covariates for 185 matched pairs from the NSW randomized experiment

Covariate	Group	Mean	25%	50%	75%
Age	Treated	25.82	20	25	29
	Control	25.70	20	25	29
Years of Education	Treated	10.35	9	11	12
	Control	10.19	9	10	11
Covariate	Group	Mean		Percent $0	
Earnings in $ in 1974	Treated	2096		71%	
	Control	2009		75%	
Earnings in $ in 1975	Treated	1532		60%	
	Control	1485		64%	
Covariate	Group	Percent			
Black	Treated	84%			
	Control	85%			
Hispanic	Treated	6%			
	Control	5%			
Married	Treated	19%			
	Control	20%			
No High School Degree	Treated	71%			
	Control	77%			

For age and years of education, means, median (50%) and quartiles (25% and 75%) are given. For earnings, the mean of all earnings (including zero earnings) and the percentage of zero earnings are given. For binary variables, the percentage is given

For the 185 matched pairs, Table 2.1 shows the distribution of the eight covariates. Before treatment, the treated and control groups looked fairly similar.

Figure 2.1 displays the earnings in 1978 after treatment for the 185 matched pairs of men. In the boxplot of 185 matched pair differences, the dotted lines are at 0 and ± 5000. It appears that treated men earned somewhat more.

In general, in this book, I will use small examples to illustrate the details of what goes on, and examples of practical size to illustrate analysis and interpretation. In that spirit, Table 2.2 is a random sample of five of the 185 pairs; it will be used to illustrate the details of randomization inference. The sample happens to consist of pairs 15, 37, 46, 151, and 181 of the 185 pairs.

2.1.2 Basic Notation

Table 2.2 exemplifies notation that will be used throughout the book. The index of the pair is i, $i = 1, 2, \ldots, 5 = I$, and the index of the person in the pair is j, $j = 1, 2$. In Table 2.2 and throughout the book, Z indicates treatment, $Z = 1$ for treated, $Z = 0$ for control, x is an observed covariate—there are eight covariates,

Fig. 2.1 Earnings in 1978, after treatment, for 185 pairs of men in the NSW randomized experiment. The dotted lines are at −\$5000, \$0, and \$5000

Table 2.2 Five pairs sampled at random from the 185 pairs in the NSW randomized experiment

id	i	j	treat Z_{ij}	age x_{ij1}	edu x_{ij2}	black x_{ij3}	hisp x_{ij4}	married x_{ij5}	nodegree x_{ij6}	re74 x_{ij7}	re75 x_{ij8}	re78 R_{ij}	Y_i
15	1	1	1	17	7	1	0	0	1	0	0	3024	1456
15	1	2	0	18	8	1	0	0	1	0	0	1568	
37	2	1	1	25	5	1	0	0	1	0	0	6182	3988
37	2	2	0	24	7	1	0	0	1	0	0	2194	
46	3	1	1	25	11	1	0	1	1	0	0	0	−45
46	3	2	0	25	11	1	0	1	1	0	0	45	
151	4	1	1	28	10	1	0	0	1	0	2837	3197	−2147
151	4	2	0	22	10	1	0	0	1	0	2175	5344	
181	5	1	1	33	12	1	0	1	0	20,280	10,941	15,953	3173
181	5	2	0	28	12	1	0	1	0	10,585	5551	12,780	

The variables are: id = pair number among the 185 pairs; pairs 1 to 5; person 1 or 2 in a pair; treat = 1 if treated, 0 if control; age in years; edu = education in years; black = 1 if black, 0 otherwise; hisp = 1 if hispanic, 0 otherwise; married = 1 if married, 0 otherwise; no degree = 1 if no high school degree, 0 otherwise; and re74, re75, and re78 are earnings in dollars in 1974, 1975, and 1978. Also, in pair i, Y_i is the treated-minus-control matched pair difference in 1978 earnings, $Y_i = (Z_{i1} - Z_{i2})(R_{i1} - R_{i2})$

x_{ijk}, $k = 1, \ldots, 8 = K$, in Table 2.2 —and R indicates a response, in this case, earnings after the end of treatment in 1978. The first pair in Table 2.2 consists of two unmarried, young black men with 7 or 8 years of education and no earnings in 1974 and 1975, before the start of treatment; after treatment, in 1978, the treated man earned more. The treated-minus-control matched pair difference in 1978 earnings

is $Y_i = (Z_{i1} - Z_{i2})(R_{i1} - R_{i2})$, so $Y_1 = (1 - 0)(3024 - 1568) = \$1456$ for the first pair of men.

It is convenient to have a symbol that represents the K covariates together, and vector notation does this, so $\mathbf{x}_{ij} = \left(x_{ij1}, \ldots, x_{ijK}\right)^T$ contains the K covariate values for the jth person or unit in the ith pair. If you are not familiar with vector notation, it will not be a problem here. Vectors and vector notation have many uses, but for the most part, in the current book they are used simply to give concise names to arrays of data. For instance, in Table 2.2, $\mathbf{x}_{11} = (17, 7, 1, 0, 0, 1, 0, 0)^T$ contains the covariate values for the first man in the first pair.

In the same way, $\mathbf{Z}$ indicates the treatment assignments for all $2I$ subjects in the I matched pairs, $\mathbf{Z} = (Z_{11}, Z_{12}, Z_{21}, \ldots, Z_{I2})^T$. For the $I = 5$ pairs in Table 2.2, $\mathbf{Z} = (1, 0, 1, 0, 1, 0, 1, 0, 1, 0)^T$. The notation is slightly redundant, because $Z_{i2} = 1 - Z_{i1}$, so that, now and then, when compact expression is needed, only the Z_{i1}'s are mentioned. Also, $\mathbf{R} = (R_{11}, R_{12}, \ldots, R_{I2})^T$ and $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_I)^T$.

It is possible to match for observed covariates, but not for a covariate that was not observed. There is an important sense in which a failure to match or control for an unobserved covariate presents no special problem in a randomized experiment but can present substantial problems in an observational study. Clarification of this distinction is one goal of the current chapter. For this purpose, it is convenient to give a name, u_{ij}, to the unobserved covariate. In the current chapter, which is focused on inference in randomized experiments, u_{ij} could be any unmeasured covariate (or any vector containing several unmeasured covariates). Success in finding and keeping a job may depend on aspects of personality, intelligence, family and personal connections, and physical appearance; then, in this chapter in connection with the NSW experiment, u_{ij} could be a vector of measurements of these attributes, a vector that was not measured by the investigators. Also, $\mathbf{u} = (u_{11}, u_{12}, \ldots, u_{I2})^T$.

To aid in reading the table, subject $j = 1$ is always the treated subject and subject $j = 2$ is always the control subject; however, strictly speaking, one should not do this. Strictly speaking, ij is the "name" of a unique person; that person had his name, ij, before treatments were randomly assigned. In a paired randomized experiment, Z_{i1} is determined by I independent flips of a fair coin, so $Z_{11} = 1$ happens with probability $\frac{1}{2}$ and $Z_{11} = 0$ happens with probability $\frac{1}{2}$; then $Z_{i2} = 1 - Z_{i1}$. That is, in Table 2.2, Z_{i1} is 1 and Z_{i2} is 0 for every i, but this would happen by chance in a paired randomized experiment with $I = 5$ pairs with probability $\left(\frac{1}{2}\right)^5 = \frac{1}{32} = 0.03125$. Strictly speaking, some of the Z_{i1}'s should be 1 and some of the Z_{i2}'s should be 1. Quantities in statistical computations, such as the treated-minus-control difference in responses, $Y_i = (Z_{i1} - Z_{i2})(R_{i1} - R_{i2})$, are unaffected by the ordering of the table. So there is a small inconsistency between the logic of random assignment and the form of a readable table. Having mentioned this once, I will adhere to the logic of random assignment in theoretical discussions, present tables in readable form, and ignore the small inconsistency.

2.2 Treatment Effects in Randomized Experiments

2.2.1 Potential Responses Under Alternative Treatments

In Table 2.2, the first man, $(i, j) = (1, 1)$, was randomly assigned to treatment, $Z_{11} = 1$, and had earnings of $R_{11} = \$3024$ in 1978, but things might have been different. Had the coin fallen differently, the first man, $(i, j) = (1, 1)$, might have been randomly assigned to control, $Z_{11} = 0$, and in this case his earnings in 1978 might have been different. We will never know what this first man's earnings would have been had the coin fallen differently, had he been assigned to control. Perhaps the treatment was completely ineffective, and perhaps this first man would have held the same job, with the same earnings, namely \$3024, had he been assigned to control. Or perhaps the treatment raised his earnings, and his earnings under control would have been lower. We will never know about this one man, but with 370 men in 185 pairs, half randomly assigned to treatment, the others to control, we can say something about what would have happened to the 370 men under treatment and under control.

Not only the first man but also each man (i, j) has two potential responses, a level of earnings he would exhibit in 1978 if assigned to treatment, r_{Tij}, and a level of earnings he would exhibit in 1978 if assigned to control, r_{Cij}. We see one of these. Specifically, if (i, j) were assigned to treatment, $Z_{ij} = 1$, we would see r_{Tij}, but if (i, j) were assigned to control, $Z_{ij} = 0$, we would see r_{Cij}. The response, R_{ij}, we actually observe from (i, j)—that is, the 1978 earnings actually recorded in Table 2.2—equal r_{Tij} if $Z_{ij} = 1$ or r_{Cij} if $Z_{ij} = 0$; that is, in a formula, $R_{ij} = Z_{ij}\, r_{Tij} + \left(1 - Z_{ij}\right) r_{Cij}$. Also, $\mathbf{r}_T = (r_{T11}, r_{T12}, \ldots, r_{TI2})^T$ and $\mathbf{r}_C = (r_{C11}, r_{C12}, \ldots, r_{CI2})^T$.

To say that the treatment has no effect on this response from (i, j) is to say $r_{Cij} = r_{Tij}$. To say that the treatment caused (i, j)'s earnings to increase by \$1000 is to say $r_{Tij} = r_{Cij} + 1000$ or $r_{Tij} - r_{Cij} = 1000$. We will never be in a position to confidently assert either of these things about a single man. It is, however, a very different thing to assert that all 370 men were unaffected by the treatment—to assert that $r_{Cij} = r_{Tij}$ for $i = 1, 2, \ldots 185$, $j = 1, 2$; in a randomized experiment, we may be in a position to confidently deny that. The hypothesis that the treatment had no effect on anyone, namely $H_0 : r_{Cij} = r_{Tij}$ for $i = 1, 2, \ldots 185$, $j = 1, 2$, is known as Ronald Fisher's [22] sharp null hypothesis of no effect. This hypothesis may be written compactly as $H_0 : \mathbf{r}_T = \mathbf{r}_C$.

The notation that expresses treatment effects as comparisons of potential responses under alternative treatments was introduced into the design of experiments by Jerzy Neyman [49] in 1923 and was used to solve various problems in randomized experiments; e.g., [13, 55, 83, 85] and [40, § 8.3]. Donald Rubin [68] first advocated use of this notation in observational studies.

2.2.2 Covariates and Outcomes

In Sect. 1.2, the distinction between covariates and outcomes was emphasized. A covariate, such as $\mathbf{x}_{ij}$ or u_{ij}, is a pretreatment quantity, so there is only one version of a covariate. A response or outcome has two potential values, $\left(r_{Tij},\, r_{Cij}\right)$, one of which is observed, namely R_{ij}, depending upon the treatment assignment Z_{ij}, that is, $R_{ij} = Z_{ij}\, r_{Tij} + \left(1 - Z_{ij}\right) r_{Cij}$.

Notice that $\left(r_{Tij},\, r_{Cij},\, \mathbf{x}_{ij},\, u_{ij}\right)$ do not change when treatments are assigned at random, that is, when Z_{ij} is determined by a coin flip or a random number, but in general the observed response, R_{ij}, does change. It is convenient to have a symbol that represents the quantities that are not changed when treatments are randomized. Let $\mathscr{F}$ denote the array of quantities $\left\{\left(r_{Tij},\, r_{Cij},\, \mathbf{x}_{ij},\, u_{ij}\right),\, i = 1, 2, \ldots, I,\; j = 1, 2\right\}$ that do not change when Z_{ij} is determined. (These quantities are fixed in Fisher's theory of randomization inference; hence, the symbol is $\mathscr{F}$.)

2.2.3 Possible Treatment Assignments and Randomization

The observed treatment assignment in Table 2.2 is $\mathbf{Z} = (1, 0, 1, 0, 1, 0, 1, 0, 1, 0)^T$, but random assignment of treatments within pairs might have picked a different assignment. Write $\mathscr{Z}$ for the set of the 2^I possible values of $\mathbf{Z}$; that is, $\mathbf{z} \in \mathscr{Z}$ if and only if $\mathbf{z} = (z_{11}, z_{12}, \ldots, z_{I2})^T$ with $z_{ij} = 0$ or $z_{ij} = 1$ and $z_{i1} + z_{i2} = 1$ for each i, j. For Table 2.2, there are $2^I = 2^5 = 32$ possible values $\mathbf{z} \in \mathscr{Z}$. Generally, if A is a finite set, then $|A|$ is the number of elements of A, so that, in particular, $|\mathscr{Z}| = 2^I$. Table 2.3 lists, in abbreviated form, the 32 possible treatment assignments for Table 2.2; the abbreviation consists in listing z_{i1} but not z_{i2} because $z_{i2} = 1 - z_{i1}$. The observed treatment assignment, $\mathbf{Z} = (1, 0, 1, 0, 1, 0, 1, 0, 1, 0)^T$ in Table 2.2, corresponds with the first row of Table 2.3. In the second row of Table 2.3, the treatment assignment for the fifth pair in Table 2.2 has been reversed, so the second man, not the first man, in that pair is assigned to treatment.

For the $I = 185$ pairs in Table 2.1, the set $\mathscr{Z}$ of possible treatment assignments contains $2^I = 2^{185} \doteq 4.9 \times 10^{55}$ possible treatment assignments, $\mathbf{z} \in \mathscr{Z}$. It would be inconvenient to list them.

What does it mean to randomly assign treatments? At an intuitive level, one assignment $\mathbf{z} \in \mathscr{Z}$ is picked at random, each having probability 2^{-I}, or $2^{-5} = \frac{1}{32} = 0.03125$ for Table 2.3. For instance, one might flip a fair coin independently five times to determine $\mathbf{Z}$ in Table 2.3. The intuition is that randomization is making a statement about how $\mathbf{Z}$ alone was determined, that is, about the marginal distribution of $\mathbf{Z}$. This intuition is not quite correct; in an important way, randomization means more than this. Specifically, in a randomized experiment, the information in $\left(r_{Tij},\, r_{Cij},\, \mathbf{x}_{ij},\, u_{ij}\right)$ (or in $\mathscr{F}$) is of no use in predicting Z_{ij}. That is, the coin is fair not just in coming up heads half the time, independently in different pairs, but

Table 2.3 The set $\mathscr{Z}$ of $32 = 2^5$ possible treatment assignments for the small version of the NSW experiment with $I = 5$ pairs

Label	Pair 1 Z_{11}	Pair 2 Z_{21}	Pair 3 Z_{31}	Pair 4 Z_{41}	Pair 5 Z_{51}
1	1	1	1	1	1
2	1	1	1	1	0
3	1	1	1	0	1
4	1	1	1	0	0
5	1	1	0	1	1
6	1	1	0	1	0
7	1	1	0	0	1
8	1	1	0	0	0
9	1	0	1	1	1
10	1	0	1	1	0
11	1	0	1	0	1
12	1	0	1	0	0
13	1	0	0	1	1
14	1	0	0	1	0
15	1	0	0	0	1
16	1	0	0	0	0
17	0	1	1	1	1
18	0	1	1	1	0
19	0	1	1	0	1
20	0	1	1	0	0
21	0	1	0	1	1
22	0	1	0	1	0
23	0	1	0	0	1
24	0	1	0	0	0
25	0	0	1	1	1
26	0	0	1	1	0
27	0	0	1	0	1
28	0	0	1	0	0
29	0	0	0	1	1
30	0	0	0	1	0
31	0	0	0	0	1
32	0	0	0	0	0

Only Z_{i1} is listed because $Z_{i2} = 1 - Z_{i1}$

more importantly the coin knows nothing about the individual and is impartial in its treatment assignments. The design of the paired randomized experiment forces $\mathbf{Z}$ to fall in $\mathscr{Z}$—that is, it forces the event $\mathbf{Z} \in \mathscr{Z}$ to occur—which is denoted concisely by saying the event $\mathscr{Z}$ occurs. Randomization in the paired randomized experiment means

$$\Pr(\mathbf{Z} = \mathbf{z} \mid \mathscr{F}, \mathscr{Z}) = \frac{1}{|\mathscr{Z}|} = \frac{1}{2^I} \quad \text{for each} \quad \mathbf{z} \in \mathscr{Z}. \tag{2.1}$$

In words, even if you knew the information in $\mathscr{F}$—that is, the information in $(r_{Tij}, r_{Cij}, \mathbf{x}_{ij}, u_{ij})$ for every person, $i = 1, 2, \ldots, I$, $j = 1, 2$—and even if you knew the structure of the paired randomized experiment, namely $\mathbf{Z} \in \mathscr{Z}$, you would be unable to use that information to predict treatment assignments because, given all of that information, the 2^I possible assignments $\mathbf{z} \in \mathscr{Z}$ have equal probabilities, 2^{-I}.

2.2.4 Interference Between Units

The notation for treatment effects appears innocuous, but it actually entails a fairly strong assumption, called "no interference between units" by David Cox [12, §2.4], namely "that the observation on one unit should be unaffected by particular assignment of treatments to other units." There would be interference between units in Table 2.2 if switching man $(i, j) = (2, 1)$—i.e., the first man in the second pair—from treatment to control would affect the earnings of the man we discussed before, $(i, j) = (1, 1)$, that is, the first man in the first pair. Interference is probably not widespread in the NSW experiment, but it cannot be entirely ruled out. For instance, a potential employer might hire two subsidized workers under NSW, but retain for unsubsidized permanent employment only the better of the two employees. In this case, switching the treatment for one man might cause another man to gain or lose a job. The two-outcomes notation makes no provision for interference between units; each unit or man has only two potential responses depending exclusively on his own treatment assignment. If there were interference between units, the response of each man would depend upon all of the treatment assignments, $(Z_{11}, Z_{12}, \ldots, Z_{I2})$. In Table 2.2, there are two ways to assign treatments in each of five pairs, so there are $32 = 2^5$ possible treatment assignments. If there were interference between units, each man would have not two potential responses but 32 potential responses, depending upon all of the treatment assignments, $(Z_{11}, Z_{12}, \ldots, Z_{52})$.

Interference between units is likely to be very limited in the NSW experiment because most men do not interfere with each other. The situation is different when the same person is studied repeatedly under different treatments in a randomized order. For instance, this is common in many experiments in cognitive neuroscience, in which a person's brain is watched using functional magnetic resonance imaging while she performs different cognitive tasks [50]. In this case, an experimental unit is one time for one person. In experiments of this kind, interference between different times for the same person is likely.

In randomized experiments with interference between units, inference ispossible but requires additional care [2, 4, 5, 20, 38, 62, 76, 78]. There is considerable simplification if only units in the same pair interfere with one another [61, §6].

2.3 Testing the Null Hypothesis of No Treatment Effect

2.3.1 Treated–Control Differences When the Null Hypothesis Is True

What Would Happen If the Null Hypothesis of No Effect Were True?

As discussed in Sect. 2.2.1, the null hypothesis of no treatment effect says each person would exhibit the same response whether assigned to treatment, $Z_{ij} = 1$, or to control, $Z_{ij} = 0$; that is, the hypothesis asserts $H_0 : r_{Cij} = r_{Tij}$ for all i, j, or concisely $H_0 : \mathbf{r}_T = \mathbf{r}_C$. In Table 2.2, if $H_0 : \mathbf{r}_T = \mathbf{r}_C$ were true, then the first man in the first pair would have earned \$3024 in 1978 whether he was assigned to treatment or to control. Similarly, if $H_0 : \mathbf{r}_T = \mathbf{r}_C$ were true, the second man in the first pair would have earned \$1568 whether he was assigned to treatment or control. This null hypothesis is not implausible; it might be true. Many treatments you or I receive during the course of a year do not change our earnings during the course of the year: we work at the same job, receive the same pay. However, whether or not someone regards H_0 as plausible has little to do with hypothesis testing—hypothesis testing is not about what goes on in someone's head, not about belief or plausibility. In hypothesis testing, we are interested in what the data say, not in what you or I might say, or believe, or find plausible. Let us leave our, perhaps conflicting, beliefs and intuitions to one side, and instead ask: What do the data say? Hypothesis testing contrasts a claim about the world to the behavior of the data: Does the behavior of the data stand in sharp contrast to the hypothesis, the claim about the world? Or would the observed data be quite ordinary were the hypothesis true? If the data from the NSW experiment would have been quite ordinary were there no effect of the treatment, then it would be hard to cite the NSW experiment as the basis for widespread use of that treatment, even if you believed in your heart of hearts that the treatment must have benefited someone. Testing the hypothesis, H_0, of no effect is the first step in a task that requires several steps; however, it is not inconsequential step, and it is a building block from which later steps are constructed, including confidence intervals and equivalence tests.

Then again, the NSW might have had a beneficial effect, raising the earnings for treated men; in this case, $H_0 : \mathbf{r}_T = \mathbf{r}_C$ would be false. A basic question is: To what extent do the data from the randomized experiment provide evidence against the null hypothesis of no treatment effect?

If the null hypothesis of no treatment effect, $H_0 : \mathbf{r}_T = \mathbf{r}_C$, were true, then depending upon the fall of the coin for the first pair in Table 2.2, the treated-minus-control difference in 1978 earnings in the first pair, namely Y_1, would be either $3024 - 1568 = 1456$ or $1568 - 3024 = -1456$. In other words, if the null hypothesis of no treatment effect were true, then the treatment does not do anything to the responses, and it affects the treated-minus-control difference in responses only by labeling one man as treated, the other as control. Recall that the observed response for man j in pair i is $R_{ij} = Z_{ij}\, r_{Tij} + \left(1 - Z_{ij}\right) r_{Cij}$ and the treated-minus-control

difference in responses in pair i is $Y_i = (Z_{i1} - Z_{i2})(R_{i1} - R_{i2})$. If the treatment had no effect, so that $r_{Tij} = r_{Cij}$ for all i, j, then $Y_i = (Z_{i1} - Z_{i2})(R_{i1} - R_{i2})$ is $Y_i = (Z_{i1} - Z_{i2})(r_{Ci1} - r_{Ci2})$ which equals $r_{Ci1} - r_{Ci2}$ if $(Z_{i1}, Z_{i2}) = (1, 0)$ or $-(r_{Ci1} - r_{Ci2})$ if $(Z_{i1}, Z_{i2}) = (0, 1)$. If the treatment has no effect, randomization just randomizes the labels "treated" and "controls"; however, it does not change anybody's earnings, and it simply changes the sign of the treated-minus-control difference in earnings.

What Experimental Results Might Occur If the Null Hypothesis of No Effect Were True?

Under the null hypothesis of no treatment effect, $H_0 : \mathbf{r}_T = \mathbf{r}_C$, Table 2.4 displays the possible treatment assignments Z_{i1} and possible treated-minus-control differences in responses Y_i for the five pairs in Table 2.2. Each row of Table 2.4 is one possible outcome of the experiment when the null hypothesis, H_0, is true. The random numbers actually gave us the first row of Table 2.4, but had the coins fallen differently, another row would have occurred. If H_0 were true, each row of Table 2.4 has probability $\frac{1}{2^5} = \frac{1}{32}$ of turning out to be the results of the experiment: flip the five coins and you get one row of Table 2.4 as your experiment. We know this without assumptions. We know that each of the 32 treatment assignments in Table 2.3 has probability $\frac{1}{2^5} = \frac{1}{32}$ because we randomized: $\mathbf{Z}$ was determined by the independent flips of five fair coins. We know "if H_0 were true, each row of Table 2.4 has probability $\frac{1}{2^5} = \frac{1}{32}$," because if H_0 were true, then different $\mathbf{Z}$'s simply change the signs of the Y_i's as in Table 2.4. Nothing is assumed.

2.3.2 The Randomization Distribution of the Mean Difference

Randomization Test of No Effect Using the Mean as the Test Statistic

In general, the null distribution of a statistic is its distribution if the null hypothesis were true. A P-value or significance level is computed with reference to a null distribution. This section considers the null distribution of the sample mean difference; it was considered by Fisher [22, Chapter 3] in the book in which he introduced randomized experimentation.

Perhaps the most familiar test statistic is the mean of the I treated-minus-control differences in response, $\overline{Y} = \frac{1}{I}\sum_{i=1}^{I} Y_i$. The null distribution of $\overline{Y}$ is the distribution of $\overline{Y}$ when the null hypothesis of no treatment effect, $H_0 : \mathbf{r}_T = \mathbf{r}_C$, is true. Using Table 2.4, we may compute $\overline{Y}$ from each of the 32 possible experimental outcomes; see Table 2.5. For instance, in the experiment as it was actually performed, $\overline{Y} = 1285$ in the first row of Table 2.5: on average, the five treated men earned \$1285 more than their matched controls. If the null hypothesis, H_0, were true, and the treatment assignment in the $i = 5$th pair were reversed, as in the second row of

Table 2.4 The set $\mathscr{Z}$ of $32 = 2^5$ possible treatment assignments and the corresponding possible treated-minus-control differences in responses $Y_i = (Z_{i1} - Z_{i2})(R_{i1} - R_{i2})$ under the null hypothesis of no treatment effect, $H_0 : \mathbf{r}_T = \mathbf{r}_C$, for the small version of the NSW experiment with $I = 5$ pairs

Label	Pair 1	Pair 2	Pair 3	Pair 4	Pair 5	Pair 1	Pair 2	Pair 3	Pair 4	Pair 5
	Z_{11}	Z_{21}	Z_{31}	Z_{41}	Z_{51}	Y_1	Y_2	Y_3	Y_4	Y_5
1	1	1	1	1	1	1456	3988	−45	−2147	3173
2	1	1	1	1	0	1456	3988	−45	−2147	−3173
3	1	1	1	0	1	1456	3988	−45	2147	3173
4	1	1	1	0	0	1456	3988	−45	2147	−3173
5	1	1	0	1	1	1456	3988	45	−2147	3173
6	1	1	0	1	0	1456	3988	45	−2147	−3173
7	1	1	0	0	1	1456	3988	45	2147	3173
8	1	1	0	0	0	1456	3988	45	2147	−3173
9	1	0	1	1	1	1456	−3988	−45	−2147	3173
10	1	0	1	1	0	1456	−3988	−45	−2147	−3173
11	1	0	1	0	1	1456	−3988	−45	2147	3173
12	1	0	1	0	0	1456	−3988	−45	2147	−3173
13	1	0	0	1	1	1456	−3988	45	−2147	3173
14	1	0	0	1	0	1456	−3988	45	−2147	−3173
15	1	0	0	0	1	1456	−3988	45	2147	3173
16	1	0	0	0	0	1456	−3988	45	2147	−3173
17	0	1	1	1	1	−1456	3988	−45	−2147	3173
18	0	1	1	1	0	−1456	3988	−45	−2147	−3173
19	0	1	1	0	1	−1456	3988	−45	2147	3173
20	0	1	1	0	0	−1456	3988	−45	2147	−3173
21	0	1	0	1	1	−1456	3988	45	−2147	3173
22	0	1	0	1	0	−1456	3988	45	−2147	−3173
23	0	1	0	0	1	−1456	3988	45	2147	3173
24	0	1	0	0	0	−1456	3988	45	2147	−3173
25	0	0	1	1	1	−1456	−3988	−45	−2147	3173
26	0	0	1	1	0	−1456	−3988	−45	−2147	−3173
27	0	0	1	0	1	−1456	−3988	−45	2147	3173
28	0	0	1	0	0	−1456	−3988	−45	2147	−3173
29	0	0	0	1	1	−1456	−3988	45	−2147	3173
30	0	0	0	1	0	−1456	−3988	45	−2147	−3173
31	0	0	0	0	1	−1456	−3988	45	2147	3173
32	0	0	0	0	0	−1456	−3988	45	2147	−3173

Only Z_{i1} is listed because $Z_{i2} = 1 - Z_{i1}$. Under the null hypothesis, different treatment assignments relabel men as treated or control, but do not change their earnings, so the sign of Y_i changes accordingly

Table 2.5 The possible treated-minus-control differences in responses $Y_i = (Z_{i1} - Z_{i2})(R_{i1} - R_{i2})$ and their mean, $\overline{Y}$, under the null hypothesis of no treatment effect, $H_0 : \mathbf{r}_T = \mathbf{r}_C$, for the small version of the NSW experiment with $I = 5$ pairs

Label	Pair 1 Y_1	Pair 2 Y_2	Pair 3 Y_3	Pair 4 Y_4	Pair 5 Y_5	Mean $\overline{Y}$
1	1456	3988	−45	−2147	3173	1285.0
2	1456	3988	−45	−2147	−3173	15.8
3	1456	3988	−45	2147	3173	2143.8
4	1456	3988	−45	2147	−3173	874.6
5	1456	3988	45	−2147	3173	1303.0
6	1456	3988	45	−2147	−3173	33.8
7	1456	3988	45	2147	3173	2161.8
8	1456	3988	45	2147	−3173	892.6
9	1456	−3988	−45	−2147	3173	−310.2
10	1456	−3988	−45	−2147	−3173	−1579.4
11	1456	−3988	−45	2147	3173	548.6
12	1456	−3988	−45	2147	−3173	−720.6
13	1456	−3988	45	−2147	3173	−292.2
14	1456	−3988	45	−2147	−3173	−1561.4
15	1456	−3988	45	2147	3173	566.6
16	1456	−3988	45	2147	−3173	−702.6
17	−1456	3988	−45	−2147	3173	702.6
18	−1456	3988	−45	−2147	−3173	−566.6
19	−1456	3988	−45	2147	3173	1561.4
20	−1456	3988	−45	2147	−3173	292.2
21	−1456	3988	45	−2147	3173	720.6
22	−1456	3988	45	−2147	−3173	−548.6
23	−1456	3988	45	2147	3173	1579.4
24	−1456	3988	45	2147	−3173	310.2
25	−1456	−3988	−45	−2147	3173	−892.6
26	−1456	−3988	−45	−2147	−3173	−2161.8
27	−1456	−3988	−45	2147	3173	−33.8
28	−1456	−3988	−45	2147	−3173	−1303.0
29	−1456	−3988	45	−2147	3173	−874.6
30	−1456	−3988	45	−2147	−3173	−2143.8
31	−1456	−3988	45	2147	3173	−15.8
32	−1456	−3988	45	2147	−3173	−1285.0

Table 2.5, then in that different randomized experiments, the mean would have been $\overline{Y} = 15.8$, or \$15.80. Similar considerations apply to all 32 possible randomized experiments.

In fact, Tables 2.5 and 2.6 give the null randomization distribution of the mean difference $\overline{Y}$. That is, the tables give the distribution of $\overline{Y}$ derived from the random assignment of treatments and the null hypothesis, H_0, that the treatment is without

Table 2.6 The randomization distribution of the sample mean $\overline{Y}$ of $I = 5$ matched pair differences in 1978 earnings under the null hypothesis of no treatment effect in the NSW experiment

y	$\Pr\left(\overline{Y} = y \mid \mathscr{F}, \mathscr{Z}\right)$	$\Pr\left(\overline{Y} \geq y \mid \mathscr{F}, \mathscr{Z}\right)$
2161.8	0.03125	0.03125
2143.8	0.03125	0.06250
1579.4	0.03125	0.09375
1561.4	0.03125	0.12500
1303.0	0.03125	0.15625
1285.0	0.03125	0.18750
892.6	0.03125	0.21875
874.6	0.03125	0.25000
720.6	0.03125	0.28125
702.6	0.03125	0.31250
566.6	0.03125	0.34375
548.6	0.03125	0.37500
310.2	0.03125	0.40625
292.2	0.03125	0.43750
33.8	0.03125	0.46875
15.8	0.03125	0.50000
−15.8	0.03125	0.53125
−33.8	0.03125	0.56250
−292.2	0.03125	0.59375
−310.2	0.03125	0.62500
−548.6	0.03125	0.65625
−566.6	0.03125	0.68750
−702.6	0.03125	0.71875
−720.6	0.03125	0.75000
−874.6	0.03125	0.78125
−892.6	0.03125	0.81250
−1285.0	0.03125	0.84375
−1303.0	0.03125	0.87500
−1561.4	0.03125	0.90625
−1579.4	0.03125	0.93750
−2143.8	0.03125	0.96875
−2161.8	0.03125	1.00000

Because the observed value of $\overline{Y}$ is \$1285, the one-sided P-value is 0.1875 and the two-sided P-value is twice that, or 0.375. With $I = 5$ pairs, the difference would be significant at the conventional 0.05 level in a one-sided test only if all five differences were positive, in which case $\overline{Y} = 2161.8$ and the one-sided P-value is 0.03125. Notice the that null distribution of $\overline{Y}$ is symmetric about zero

effect. Table 2.5 lists the results in the order of Table 2.3, while Table 2.6 sorts Table 2.5 into increasing order of the possible values of $\overline{Y}$. It is the same distribution, but Table 2.6 is easier to read. In both tables, each possible outcome or row has probability $\frac{1}{2^5} = \frac{1}{32} = 0.03125$. The tail probability, $\Pr\left(\overline{Y} \geq y \middle| \mathscr{F}, \mathscr{Z}\right)$, used in computing significance levels or P-values, is also given in Table 2.6. The one-sided P-value for testing no treatment effect in Table 2.2 is $\Pr\left(\overline{Y} \geq 1285 \middle| \mathscr{F}, \mathscr{Z}\right) = 0.1875 = \frac{6}{32}$ because 6 of the 32 random assignments produce means $\overline{Y}$ of 1285 or more when the null hypothesis of no effect is true. The two-sided P-value is twice this one-sided P-value,[2] or $2 \times \Pr\left(\overline{Y} \geq 1285 \middle| \mathscr{F}, \mathscr{Z}\right) = 0.375 = \frac{12}{32}$, which equals $\Pr\left(\overline{Y} \leq -1285 \middle| \mathscr{F}, \mathscr{Z}\right) + \Pr\left(\overline{Y} \geq 1285 \middle| \mathscr{F}, \mathscr{Z}\right)$ for the distribution in Table 2.6 because the distribution is symmetric about zero.

The Reasoned Basis for Inference in Experiments

Fisher [22, Chapter 2] spoke of randomization as the "reasoned basis" for causal inference in experiments. By this, he meant that the distribution in Table 2.6 for the mean difference $\overline{Y}$ and similar randomization distributions for other test statistics provided a valid test of the hypothesis of no effect caused by the treatment. Moreover, these tests required no assumptions whatsoever. Fisher placed great emphasis on this, so it is worthwhile to consider what he meant and why he regarded the issue as important.

The distribution in Table 2.6 was derived from two considerations. The first consideration is that the experimenter used coins or random numbers to assign treatments, picking one of the 32 treatment assignments in Table 2.3, each with probability 1/32. Provided the experimenter is not dishonest in describing the experiment, there is no basis for doubting this first consideration. The first consideration is not an assumption; it is a fact describing how the experiment was conducted. The second consideration is the null hypothesis of no treatment effect. Table 2.6 is the null distribution of the mean difference, $\overline{Y}$; that is, Table 2.6 is the distribution $\overline{Y}$ *would* have if the null hypothesis of no treatment effect *were* true. The null hypothesis may well be false—perhaps the experiment was undertaken in the hope of showing it is false—but that is beside the point because Table 2.6 is the distribution $\overline{Y}$ *would* have if the null hypothesis of no treatment effect *were* true. One tests a null hypothesis using $\overline{Y}$ by contrasting the actual behavior of $\overline{Y}$ with the

[2] In general, if you want a two-sided P-value, compute both one-sided P-values, double the smaller one, and take the minimum of this value and 1. This approach views the two-sided P-value as a correction for testing twice [14]. Both the sample mean in the current section and Wilcoxon's signed rank statistic in the next section have randomization distributions under the null hypothesis that are symmetric, and with a symmetric null distribution there is little ambiguity about the meaning of a two-sided P-value. When the null distribution is not symmetric, different definitions of a two-sided P-value can give slightly different answers. As discussed in [14], the view of a two-sided P-value as a correction for testing twice is one sensible approach in all cases. For related results, see [73].

behavior $\overline{Y}$ would have if the null hypothesis were true. A small P-value indicates that the behavior $\overline{Y}$ actually exhibited would have been remarkably improbable if the null hypothesis were true. In this sense too, the second consideration, the null hypothesis, is not an assumption. One does not assume or believe the null hypothesis to be true when testing it; rather, certain logical and probabilistic consequences of the null hypothesis are worked out for the purpose of evaluating whether the hypothesis and the observed data are markedly incompatible. In brief, nothing was assumed to be true in working out the distribution in Table 2.6, that is, in testing the hypothesis that the treatment caused no effect.

You sometimes hear it said that "You cannot prove causality with statistics." One of my professors, Frederick Mosteller, would often say: "You can only prove causality with statistics." When he said this, he was referring to randomization as the reasoned basis for causal inference in experiments.

Some History: Random Assignment and Normal Errors

Concerning Table 2.6, one might reasonably ask how this null randomization distribution of the mean treated-minus-control difference, $\overline{Y}$, compares with the t-distribution that might be used if the Y_i were assumed to be Normally distributed with constant variance. The null distribution in Table 2.6 made no assumption that the Y_i were Normal with constant variance; rather, it acknowledged that the experimenter did indeed randomize treatment assignments. Questions of this sort received close attention in early work on randomization inference; see, for instance, [83]. This attention focused on the moments of the distributions and their limiting behavior as the number, I, of pairs increases. Generally, the two approaches to inference gave similar inferences for large I when the $\left(r_{Cij}, r_{Tij}\right)$ were sufficiently well-behaved, but could diverge in other situations. In early work [22, 83], this was often understood to mean that randomization formed a justification for inferences based on Normal models without requiring the assumptions of those models to be true. However, it was not long before attention focused on the divergences between randomization inference and Normal theory, so the concern ceased to be with justifying Normal theory when applied to well-behaved data from a randomized experiment [21]; instead, concern shifted to inferences that would be valid and efficient whether or not the data are well-behaved.[3] This leads to the randomization distribution of statistics other than the mean difference, $\overline{Y}$, because the mean was found to be one of the least efficient statistics for responses that are not so well-behaved [1]. Two such statistics will be discussed: Frank Wilcoxon's signed rank

[3] I write "Normal" distribution rather than "normal" distribution because Normal is the name, not the description, of a distribution, in the same way that "Sitting Bull" is the name of a man, not a description of that man. The exponential distribution is exponential, the logistic distribution is logistic, but the Normal distribution is not normal. Under normal circumstances, data are not Normally distributed.

statistic [84] in Sect. 2.3.3 and the randomization distribution of the statistics used in Peter Huber's m-estimation [37, 47] in Sect. 2.8. The close connection between rank statistics and m-estimates is developed in detail in [39]. Sections 2.3.3 and 2.8 play different roles in this book, and it is not necessary to read Sect. 2.8. The signed rank statistic will be discussed at many points in the book. In contrast, m-estimation is mentioned here to emphasize that the basic issues discussed here and later are not tied to any one statistic and may be developed for large classes of statistics in an entirely parallel way.

2.3.3 The Randomization Distribution of Wilcoxon's Statistic

Computing the Randomization Distribution Under the Null Hypothesis of No Effect

Among robust alternatives to the mean treated-minus-control difference, $\overline{Y}$, by far the most popular in practice is Wilcoxon's signed rank statistic, T; see [84] or [43, §3.2]. As with $\overline{Y}$, Wilcoxon's statistic uses the treated-minus-control differences in responses, $Y_i = (Z_{i1} - Z_{i2})(R_{i1} - R_{i2})$, in the I pairs. First, the absolute differences $|Y_i|$ are calculated; then these absolute differences are assigned ranks q_i from smallest to largest, from 1 to I; then Wilcoxon's signed rank statistic, T, is the sum of those ranks q_i for which Y_i is positive. With $Y_1 = 1456$, $Y_2 = 3988$, $Y_3 = -45$, $Y_4 = -2147$, $Y_5 = 3173$ in Table 2.2, the absolute differences are $|Y_1| = 1456$, $|Y_2| = 3988$, $|Y_3| = 45$, $|Y_4| = 2147$, $|Y_5| = 3173$, with ranks $q_1 = 2$, $q_2 = 5$, $q_3 = 1$, $q_4 = 3$, and $q_5 = 4$, and Y_1, Y_2, Y_5 are positive, so $T = q_1 + q_2 + q_5 = 2 + 5 + 4 = 11$.

In general, Wilcoxon's signed rank statistic is $T = \sum_{i=1}^{I} \text{sgn}(Y_i) \cdot q_i$, where $\text{sgn}(a) = 1$ if $a > 0$, $\text{sgn}(a) = 0$ if $a \leq 0$, and q_i is the rank of $|Y_i|$.

As in Sect. 2.3.2 and Table 2.5, if the null hypothesis of no treatment effect, $H_0 : \mathbf{r}_T = \mathbf{r}_C$, were true then $Y_i = (Z_{i1} - Z_{i2})(r_{Ci1} - r_{Ci2})$, and randomization simply changes the signs of the Y_i, as in Table 2.4. For each row of Table 2.4, Wilcoxon's signed rank statistic, T, may be computed, and this is done in Table 2.7. Under the null hypothesis of no treatment effect, $H_0 : \mathbf{r}_T = \mathbf{r}_C$, each row of Table 2.7 has a probability of $\frac{1}{2^5} = \frac{1}{32}$ of becoming the observed experimental results in the five pairs, and 7 rows produce values of T that are 11 or more, so the one-sided P-value is $\Pr(T \geq 11 | \mathscr{F}, \mathscr{Z}) = \frac{7}{32} = 0.21875$. Table 2.8 reorganizes Table 2.7, sorting by the possible values of the test statistic and removing repetitions of values that occur several times. In parallel with Sect. 2.3.2, the two-sided P-value is twice this one-sided P-value[4] or $2 \times \Pr(T \geq 11 | \mathscr{F}, \mathscr{Z}) = \frac{14}{32} = 0.4375$ which equals

[4]In practice, statistical software is used to do the calculations. In the statistical package R, the command wilcox.test(.) is used. Consider, first, the $I = 5$ matched pair differences in 1978 earnings in Table 2.7, which are contained in the vector `dif5`.

```
> dif5
```

Table 2.7 The possible treated-minus-control differences in responses $Y_i = (Z_{i1} - Z_{i2})(R_{i1} - R_{i2})$ and Wilcoxon's signed rank statistic, T, under the null hypothesis of no treatment effect, $H_0 : \mathbf{r}_T = \mathbf{r}_C$, for the small version of the NSW experiment with $I = 5$ pairs

Label	Pair 1	Pair 2	Pair 3	Pair 4	Pair 5	Wilcoxon's
	Y_1	Y_2	Y_3	Y_4	Y_5	T
1	1456	3988	−45	−2147	3173	11
2	1456	3988	−45	−2147	−3173	7
3	1456	3988	−45	2147	3173	14
4	1456	3988	−45	2147	−3173	10
5	1456	3988	45	−2147	3173	12
6	1456	3988	45	−2147	−3173	8
7	1456	3988	45	2147	3173	15
8	1456	3988	45	2147	−3173	11
9	1456	−3988	−45	−2147	3173	6
10	1456	−3988	−45	−2147	−3173	2
11	1456	−3988	−45	2147	3173	9
12	1456	−3988	−45	2147	−3173	5
13	1456	−3988	45	−2147	3173	7
14	1456	−3988	45	−2147	−3173	3
15	1456	−3988	45	2147	3173	10
16	1456	−3988	45	2147	−3173	6
17	−1456	3988	−45	−2147	3173	9
18	−1456	3988	−45	−2147	−3173	5
19	−1456	3988	−45	2147	3173	12
20	−1456	3988	−45	2147	−3173	8
21	−1456	3988	45	−2147	3173	10
22	−1456	3988	45	−2147	−3173	6
23	−1456	3988	45	2147	3173	13
24	−1456	3988	45	2147	−3173	9
25	−1456	−3988	−45	−2147	3173	4
26	−1456	−3988	−45	−2147	−3173	0
27	−1456	−3988	−45	2147	3173	7
28	−1456	−3988	−45	2147	−3173	3
29	−1456	−3988	45	−2147	3173	5
30	−1456	−3988	45	−2147	−3173	1
31	−1456	−3988	45	2147	3173	8
32	−1456	−3988	45	2147	−3173	4
All	Absolute Differences					
1–32	1456	3988	45	2147	3173	
All	Ranks of Absolute Differences					
1–32	2	5	1	3	4	

Table 2.8 The randomization distribution of Wilcoxon's signed rank statistic, T, for $I = 5$ matched pair differences in 1978 earnings under the null hypothesis of no treatment effect in the NSW experiment

t	Count	$\Pr(T = t \mid \mathscr{F}, \mathscr{Z})$	$\Pr(T \geq t \mid \mathscr{F}, \mathscr{Z})$
15	1	0.03125	0.03125
14	1	0.03125	0.06250
13	1	0.03125	0.09375
12	2	0.06250	0.15625
11	2	0.06250	0.21875
10	3	0.09375	0.31250
9	3	0.09375	0.40625
8	3	0.09375	0.50000
7	3	0.09375	0.59375
6	3	0.09375	0.68750
5	3	0.09375	0.78125
4	2	0.06250	0.84375
3	2	0.06250	0.90625
2	1	0.03125	0.93750
1	1	0.03125	0.96875
0	1	0.03125	1.00000

Count is the number of treatment assignments that yield $T = t$; for instance, there are two ways to obtain $T = 12$. Because the observed value of T is 11, the one-sided P-value is 0.21875 and the two-sided P-value is twice that, or 0.4375. With $I = 5$ pairs, the difference would be significant at the conventional 0.05 level in a one-sided test only if all five differences were positive, in which case $T = 15$ and the one-sided P-value is 0.03125. Notice the that null distribution of T is symmetric about its expectation, $I(I+1)/4 = 5(5+1)/4 = 7.5$

$\Pr(T \leq 4 \mid \mathscr{F}, \mathscr{Z}) + \Pr(T \geq 11 \mid \mathscr{F}, \mathscr{Z})$ for the distribution in Table 2.7 because the distribution is symmetric about its expectation $I(I+1)/4 = 5(5+1)/4 = 7.5$.

The null distribution of Wilcoxon's T is, in many ways, a very simple distribution. In Table 2.7, as the signs change, the absolute values $|Y_i|$ stay the same, so the ranks of the absolute values also stay the same; see the bottom of Table 2.7. That is, if $H_0 : \mathbf{r}_T = \mathbf{r}_C$, were true, then the $|Y_i| = |r_{Ci1} - r_{Ci2}|$ and q_i are fixed (i.e., determined by $\mathscr{F}$) as the treatment assignment $\mathbf{Z}$ varies, so Wilcoxon's T is changing only as the I signs change, and the I signs are

```
[1] 1456 3988 -45 -2147 3173
> wilcox.test(dif5)
Wilcoxon signed rank test
data: dif5
V = 11, p-value = 0.4375
```

In the case of $I = 5$, `wilcox.test` performs calculations that exactly reproduce the value $T = 11$ and the two-sided P-value of 0.4375. By adjusting the call to `wilcox.test`, either of the two one-sided P-values may be obtained.

independent. In a formula, under the null hypothesis of no effect, the signed rank statistic is $T = \sum_{i=1}^{I} \operatorname{sgn}\{(Z_{i1} - Z_{i2})(r_{Ci1} - r_{Ci2})\} \cdot q_i$ where $\operatorname{sgn}(a) = 1$ if $a > 0$, $\operatorname{sgn}(a) = 0$ if $a \leq 0$, and q_i is the rank of $|r_{Ci1} - r_{Ci2}|$. The expectation $E(T|\mathscr{F}, \mathscr{Z})$ and variance $\operatorname{var}(T|\mathscr{F}, \mathscr{Z})$ of the null distribution in Table 2.8 have simple formulas. Write $s_i = 1$ if $|Y_i| > 0$ and $s_i = 0$ if $|Y_i| = 0$. If $H_0 : \mathbf{r}_T = \mathbf{r}_C$ were true in a randomized experiment, then $E(T|\mathscr{F}, \mathscr{Z}) = (1/2)\sum_{i=1}^{I} s_i\, q_i$ and $\operatorname{var}(T|\mathscr{F}, \mathscr{Z}) = (1/4)\sum_{i=1}^{I} (s_i\, q_i)^2$. When there are no ties, these formulas simplify because the ranks are then $1, 2, \ldots, I$. If there are no ties, in the sense that $|Y_i| > 0$ for all i and all of the $|Y_i|$ are distinct, then the formulas simplify to $E(T|\mathscr{F}, \mathscr{Z}) = I(I+1)/4$ and $\operatorname{var}(T|\mathscr{F}, \mathscr{Z}) = I(I+1)(2I+1)/24$; see [43, §3.2] Moreover, under very mild conditions, as $I \to \infty$, the null distribution of $\{T - E(T|\mathscr{F}, \mathscr{Z})\}/\sqrt{\operatorname{var}(T|\mathscr{F}, \mathscr{Z})}$ converges in distribution to the standard Normal distribution [43].

Testing No Effect with All $I = 185$ Matched Pairs

If Wilcoxon's test is applied to all $I = 185$ matched pairs in Fig. 2.1, then the two-sided P-value for testing no effect is 0.009934. When all $I = 185$ pairs are considered, the difference in 1978 earnings in Fig. 2.1 appears to be an effect caused by the intervention. The difference in 1978 earnings is too large and systematic to be due to chance, the flip of the coin that assigned one man to treatment, the next to control.[5]

The procedure used in Table 2.5 for the mean, $\overline{Y}$, and in Table 2.7 for Wilcoxon's statistic, T, are entirely parallel and work in a similar way for any test statistic.[6]

[5]Performing the calculations in R is straightforward. The $I = 185$ paired differences in 1978 earnings are contained in the vector `dif`.

```
> length(dif)
[1] 185
> dif[1:6]
[1] 3889.711 -4733.929 18939.192 7506.146 -6862.342 -8841.886
> wilcox.test(dif)
 Wilcoxon signed rank test with continuity correction
data: dif
V = 9025, p-value = 0.009934
alternative hypothesis: true location is not equal to 0
```

Now, with $I = 185$ pairs, the statistic is $T = 9025$ and the two-sided P-value is 0.009934. In computing this P-value, R has taken appropriate account of ties of both kinds and has made use of a Normal approximation to the null distribution of T. It has also employed a "continuity correction" intended to improve the accuracy of approximate P-values obtained from the Normal distribution.

[6]This statement is true conceptually. That is, the statement would be true if we had an infinitely fast computer. As a practical matter, producing a table like Table 2.5 for large I is quite tedious, even for a computer, so randomization distributions are obtained either with clever algorithms [51] or with the aid of large sample approximations [43]. It is not necessary to focus on these technical details in a chapter whose focus is on concepts. In understanding concepts, we can pretend for the moment that we have an infinitely fast computer.

The null distribution of Wilcoxon's T in Table 2.8 has a curious property not shared by the null distribution of the mean difference, $\overline{Y}$, in Table 2.6. The null distribution of $\overline{Y}$ in Table 2.6 depends on the numeric values of Y_i, so this distribution could not have been written down before the experiment was conducted. You need to have the Y_i's to construct Table 2.6. In contrast, the distribution of T in Table 2.8 does not depend upon the differences in earnings, the Y_i's, provided the $|Y_i|$ are nonzero and different (or untied) so that they can be ranked 1, 2, 3, 4, 5 and each Y_i is either positive or negative. Provided the $I = 5$ earnings differences are untied, the null distribution in Table 2.8 could be written down before the experiment is conducted, and indeed the distribution in Table 2.8 appears at the back of many textbooks.

Ties and Distribution-Free Statistics

What happens to Wilcoxon's T when some $|Y_i|$ are tied? If several $|Y_i|$ are equal, it does not seem reasonable to assign them different ranks, so instead they are assigned the average of the ranks they would have had if they had differed ever so slightly. If $|Y_4| = |Y_5|$ and they would have had ranks 3 and 4 had they not been tied, then they are both assigned rank 3.5. Under the null hypothesis of no effect, if $|Y_i| = 0$, then $Y_i = 0$ no matter how treatments are assigned, and pair i contributes 0 to Wilcoxon's T for every treatment assignment. With ties, the null distribution of T is determined as has been done twice before in Tables 2.5 and 2.7; however, the null distribution now depends not only on the sample size I but also on the specific pattern of ties, so the distribution is not known in advance of the experiment.

The phrase "distribution-free statistic" is used in a variety of related but slightly distinct senses. In one of these senses, both the sample mean, $\overline{Y}$, and Wilcoxon's signed rank statistic, T, always have randomization distributions, as computed in Tables 2.5 and 2.7, and if there are no ties among the $|Y_i|$, then T is distribution free, in the sense that its null distribution does not depend upon the values of the Y_i.

2.4 Testing Other Hypotheses; Confidence Intervals; Point Estimates

2.4.1 Testing a Constant, Additive Treatment Effect

Section 2.3 tested Fisher's [22] sharp null hypothesis of no treatment effect, which asserts that each subject would exhibit the same response whether assigned to treatment or control, that is, $H_0 : r_{Tij} = r_{Cij}$ for $i = 1, \ldots, I$, $j = 1, 2$, or equivalently $H_0 : \mathbf{r}_T = \mathbf{r}_C$. This test for no effect quickly provides the basis for testing any hypothesis that specifies an effect, and from this to confidence intervals.

The simplest case is considered first, namely a constant, additive treatment effect, τ, in which $r_{Tij} = r_{Cij} + \tau$ for $i = 1, \ldots, I$, $j = 1, 2$, or equivalently $\mathbf{r}_T = \mathbf{r}_C + \mathbf{1}\tau$, where $\mathbf{1} = (1, 1, \ldots, 1)^T$. In this case, the observed response, $R_{ij} = Z_{ij}\, r_{Tij} + \left(1 - Z_{ij}\right) r_{Cij}$ from the jth subject in pair i becomes $R_{ij} = Z_{ij}\left(r_{Cij} + \tau\right) + \left(1 - Z_{ij}\right) r_{Cij} = r_{Cij} + \tau Z_{ij}$, that is, the observed response, R_{ij}, equals the response to control, r_{Cij}, plus the treatment effect τ if ij is assigned to treatment, $Z_{ij} = 1$. Of course, R_{ij} and Z_{ij} are observed, but r_{Cij} and τ are not.

The hypothesis of a constant, additive effect is not strictly plausible for any $\tau \neq 0$ in the NSW study in Sect. 2.1. For instance, in Table 2.2, the $j = 1$st man in the $i = 3$rd pair was treated but earned \$0.00 in 1978. Because earnings are nonnegative, the only nonnegative constant effect τ that can yield \$0.00 for a treated man is $\tau = 0$. Indeed, in Fig. 2.1, a quarter of the controls had \$0.00 of earnings in 1978. Moreover, even if negative earnings were possible, the boxplots in Fig. 2.1 for all 185 pairs of men strongly suggest the treatment effect is not an additive constant. In a very large randomized experiment with an additive, constant effect τ, the two boxplots of responses, R_{ij}, on the left in Fig. 2.1 would look the same, except that the treated boxplot would be shifted up by τ, and the boxplot of pair differences, Y_i, on the right in Fig. 2.1 would be symmetric about τ. In fact, neither of these two conditions appears to be true in Fig. 2.1. After illustrating the most commonly used methods for an additive effect, three alternative approaches will be considered for the NSW study: a multiplicative effect, a truncated additive effect similar to that found in a Tobit model, and "attributable effects" of various kinds. As will be seen, although the hypothesized effects are different, the method of inference is much the same.

Suppose that we hypothesized that the additive effect τ is some particular number, τ_0. For instance, in Sect. 2.1, a rather extreme hypothesis says that the NSW program increases earnings by \$5000, or $H_0 : \tau = 5000$, or $H_0 : \tau = \tau_0$ with $\tau_0 = 5000$. It will soon be seen that this hypothesis is quite incompatible with the data for $I = 185$ pairs. If this hypothesis were true, the observed response would be $R_{ij} = r_{Cij} + \tau_0\, Z_{ij} = r_{Cij} + 5000 Z_{ij}$, so r_{Cij} could be computed from the observed $\left(R_{ij}, Z_{ij}\right)$ together with the supposedly true hypothesis, $H_0 : \tau = 5000$, as $r_{Cij} = R_{ij} - \tau_0\, Z_{ij} = R_{ij} - 5000 Z_{ij}$. Moreover, the matched treated-minus-control difference in observed earnings is $Y_i = (Z_{i1} - Z_{i2})\,(R_{i1} - R_{i2})$, so if the hypothesis $H_0 : \tau = \tau_0$ were true,

$$
\begin{aligned}
Y_i &= (Z_{i1} - Z_{i2})\,(R_{i1} - R_{i2}) \\
&= (Z_{i1} - Z_{i2})\,\{(r_{Ci1} + \tau_0\, Z_{i1}) - (r_{Ci2} + \tau_0\, Z_{i2})\} \\
&= \tau_0 + (Z_{i1} - Z_{i2})\,(r_{Ci1} - r_{Ci2})\,,
\end{aligned}
$$

and $Y_i - \tau_0 = (Z_{i1} - Z_{i2})\,(r_{Ci1} - r_{Ci2})$. In other words, if $H_0 : \tau = \tau_0$ were true with $\tau_0 = 5000$, then the observed difference in earnings, Y_i, in pair i, less $\tau_0 = 5000$, would equal $\pm\,(r_{Ci1} - r_{Ci2})$ depending upon the random treatment assignment $Z_{i1} - Z_{i2}$ in pair i. In particular, if $H_0 : \tau = 5000$ were true, then the sign of $Y_i - 5000$ would be determined by a coin flip, and we would expect about

Table 2.9 Five matched pair differences from the NSW experiment adjusted for the hypothesis $H_0 : \tau = 5000$

i	1	2	3	4	5
Y_i	1456	3988	−45	−2147	3173
$\tau_0 = 5000$	5000	5000	5000	5000	5000
$Y_i - \tau_0$	−3544	−1012	−5045	−7147	−1827

half of the $Y_i - 5000$ to be positive, half to be negative. If $H_0 : \tau = 5000$ were true for the five pairs in Table 2.2, then, for pair $i = 1$, the adjusted difference is $r_{C11} - r_{C12} = Y_1 - 5000 = 1456 - 5000 = -3544$. The five adjusted differences are listed in Table 2.9. Even with just the five pairs in Table 2.9, the hypothesis $H_0 : \tau = 5000$ does not look plausible: all five adjusted differences are negative, whereas if the hypothesis were true, the signs would be ± 1 independently with probability $\frac{1}{2}$. Only one of the 32 treatment assignments $\mathbf{z} \in \mathscr{Z}$ would give negative signs to all five adjusted responses, and that one assignment has probability $1/32 = 0.03125$ in a randomized experiment. If $H_0 : \tau = 5000$ were true, to produce Table 2.9 would require quite a bit of bad luck when $\mathbf{Z}$ is picked at random from $\mathscr{Z}$.

Table 2.10 displays all 32 treatment assignments, $\mathbf{z} \in \mathscr{Z}$, together with all possible adjusted responses, $Y_i - \tau_0 = (Z_{i1} - Z_{i2})\,(r_{Ci1} - r_{Ci2})$, when $H_0 : \tau = 5000$ is true, together with Wilcoxon's signed rank statistic T computed from these adjusted responses. In the first row, for the actual treatment assignment, $T = 0$ because all five differences are negative. From Table 2.10, $\Pr(T \le 0 | \mathscr{F}, \mathscr{Z}) = 1/32 = 0.03125$ if $H_0 : \tau = 5000$ is true, and the two-sided P-value is $\Pr(T \le 0 | \mathscr{F}, \mathscr{Z}) + \Pr(T \ge 15 | \mathscr{F}, \mathscr{Z}) = 2/32 = 0.0625$.

Repeating the test of $H_0 : \tau = 5000$ for all $I = 185$ pairs yields a two-sided P-value of 2.8×10^{-8}. An effect of $\tau = 5000$ dollars is too large to be compatible with the observed data.

2.4.2 *Confidence Intervals for a Constant, Additive Effect*

If the treatment has an additive effect, $r_{Tij} = r_{Cij} + \tau$ for $i = 1, \ldots, I$, $j = 1, 2$, then a 95% confidence set for the additive treatment effect, τ, is formed by testing each hypothesis $H_0 : \tau = \tau_0$ and retaining for the confidence set the values of τ_0 not rejected at the 5% level. In general, a $1 - \alpha$ confidence set is the set of hypothesized values of a parameter not rejected by a level α test [44, §3.5] . If the parameter is a number, such as τ, then a $1-\alpha$ confidence interval is the shortest interval containing the $1 - \alpha$ confidence set, and this interval may be a half-line or the entire line. For many simple statistics, including Wilcoxon's signed rank statistic, the confidence set and the confidence interval for τ are the same [43, Chapter 4].

The five pairs in Table 2.10 cannot reject at the 5% level any two-sided hypothesis using Wilcoxon's statistic, T, because the smallest two-sided significance level is $\Pr(T \le 0 | \mathscr{F}, \mathscr{Z}) + \Pr(T \ge 15 | \mathscr{F}, \mathscr{Z}) = 2/32 = 0.0625$. However, a $1-0.0625 = 93.75\%$ confidence interval for τ may be constructed. The endpoints of the confidence interval are determined in Table 2.11, where Wilcoxon's signed rank

Table 2.10 Using Wilcoxon's signed rank statistic T to test the null hypothesis that the training program increased wages by an additive constant of \$5000, that is, $H_0 : r_{Tij} = r_{Cij} + 5000$ for all i, j

Label	Pair 1	Pair 2	Pair 3	Pair 4	Pair 5	Pair 1	Pair 2	Pair 3	Pair 4	Pair 5	
	Z_{11}	Z_{21}	Z_{31}	Z_{41}	Z_{51}	$Y_1 - \tau_0$	$Y_2 - \tau_0$	$Y_3 - \tau_0$	$Y_4 - \tau_0$	$Y_5 - \tau_0$	T
1	1	1	1	1	1	−3544	−1012	−5045	−7147	−1827	0
2	1	1	1	1	0	−3544	−1012	−5045	−7147	1827	2
3	1	1	1	0	1	−3544	−1012	−5045	7147	−1827	5
4	1	1	1	0	0	−3544	−1012	−5045	7147	1827	7
5	1	1	0	1	1	−3544	−1012	5045	−7147	−1827	4
6	1	1	0	1	0	−3544	−1012	5045	−7147	1827	6
7	1	1	0	0	1	−3544	−1012	5045	7147	−1827	9
8	1	1	0	0	0	−3544	−1012	5045	7147	1827	11
9	1	0	1	1	1	−3544	1012	−5045	−7147	−1827	1
10	1	0	1	1	0	−3544	1012	−5045	−7147	1827	3
11	1	0	1	0	1	−3544	1012	−5045	7147	−1827	6
12	1	0	1	0	0	−3544	1012	−5045	7147	1827	8
13	1	0	0	1	1	−3544	1012	5045	−7147	−1827	5
14	1	0	0	1	0	−3544	1012	5045	−7147	1827	7
15	1	0	0	0	1	−3544	1012	5045	7147	−1827	10
16	1	0	0	0	0	−3544	1012	5045	7147	1827	12
17	0	1	1	1	1	3544	−1012	−5045	−7147	−1827	3
18	0	1	1	1	0	3544	−1012	−5045	−7147	1827	5
19	0	1	1	0	1	3544	−1012	−5045	7147	−1827	8
20	0	1	1	0	0	3544	−1012	−5045	7147	1827	10
21	0	1	0	1	1	3544	−1012	5045	−7147	−1827	7
22	0	1	0	1	0	3544	−1012	5045	−7147	1827	9
23	0	1	0	0	1	3544	−1012	5045	7147	−1827	12
24	0	1	0	0	0	3544	−1012	5045	7147	1827	14
25	0	0	1	1	1	3544	1012	−5045	−7147	−1827	4
26	0	0	1	1	0	3544	1012	−5045	−7147	1827	6
27	0	0	1	0	1	3544	1012	−5045	7147	−1827	9
28	0	0	1	0	0	3544	1012	−5045	7147	1827	11
29	0	0	0	1	1	3544	1012	5045	−7147	−1827	8
30	0	0	0	1	0	3544	1012	5045	−7147	1827	10
31	0	0	0	0	1	3544	1012	5045	7147	−1827	13
32	0	0	0	0	0	3544	1012	5045	7147	1827	15

Table 2.11 The endpoints of the 93.75% confidence interval for an additive treatment effect τ in five pairs from the NSW experiment

τ_0	−2147.0001	−2147	3987.9999	3988
T	15	14	1	0
P-value	.0625	0.1250	0.1250	.0625

Wilcoxon's signed rank statistic T takes steps at −2147 and 3988, which define the endpoints of the two-sided confidence interval [−2147, 3988]

statistic, T, rejects hypotheses below $H_0 : \tau = -2147$ and above $H_0 : \tau = 3988$ in a two-sided, 0.0625 level test, and the 93.75% confidence interval for τ is [−2147, 3988].

With the full sample of $I = 185$ pairs, the 95% confidence interval for τ is [391, 2893] or between \$391 and \$2893, a long interval but one that excludes zero treatment effect.[7]

2.4.3 Hodges–Lehmann Point Estimates of Effect

In Sect. 2.4.1, the hypothesis of an additive treatment effect, $H_0 : r_{Tij} = r_{Cij} + \tau_0$ for $i = 1, \ldots, I$, $j = 1, 2$, was tested by applying Wilcoxon's signed rank statistic to the adjusted pair differences, $Y_i - \tau_0$, which equal $(Z_{i1} - Z_{i2})\,(r_{Ci1} - r_{Ci2})$ if H_0 is true. In Sect. 2.4.2, a $1 - \alpha$ confidence interval for an additive effect was formed by testing every value τ_0 and retaining for the interval all values not rejected by the Wilcoxon test at level α. The same logic was used by Joseph Hodges and Erich Lehmann [33, 43] to create a point estimate. They asked: What value τ_0 when subtracted from Y_i makes Wilcoxon's statistic T computed from $Y_i - \tau_0$

[7]Continuing the discussion from footnote 5, calculating the confidence interval in R is straightforward: the call to `wilcox.test(.)` includes the option `conf.int=T`.

```
 >wilcox.test(dif,conf.int=T)
 Wilcoxon signed rank test with continuity correction
data: dif
V = 9025, p-value = 0.009934
alternative hypothesis: true location is not equal to 0
95 percent confidence interval:
 391.359 2893.225
sample estimates:
(pseudo)median
 1639.383
```

In addition to the confidence interval, R supplies the Hodges–Lehmann point estimate 1639.383 of τ, as discussed in Sect. 2.4.3. In the R output, the Hodges–Lehmann estimate is labeled a (pseudo) median.

In calculating this confidence interval, R makes use of various computational shortcuts. Because these computational shortcuts do not change the confidence interval, because they make the subject seem more complex than it is, and because they do not generalize to other situations, I do not discuss such shortcuts in this book.

behave the way we expect Wilcoxon's statistic to behave with a true hypothesis? As noted in Sect. 2.3.3, when Wilcoxon's signed rank statistic T is computed from $(Z_{i1} - Z_{i2})(r_{Ci1} - r_{Ci2})$, it has expectation $E(T|\mathscr{F}, \mathscr{Z}) = I(I+1)/4$ provided $|r_{Ci1} - r_{Ci2}| > 0$ for all i. Intuitively, one finds a value, $\widehat{\tau}$, such that Wilcoxon's statistic T equals $I(I+1)/4$ when T is computed from $Y_i - \widehat{\tau}$. The intuition does not quite work, because the Wilcoxon statistic is discrete, taking on a finite number of values, so it may never equal $I(I+1)/4$ or it may equal $I(I+1)/4$ for an interval of values of τ_0. So Hodges and Lehmann added two very small patches. If the signed rank statistic never equals $I(I+1)/4$, then there is a unique value of τ_0 where T passes $I(I+1)/4$, and that unique value is the Hodges–Lehmann estimate, $\widehat{\tau}$. If the signed rank statistic equals $I(I+1)/4$ for an interval of values of τ_0, then the midpoint of that interval is the Hodges–Lehmann estimate, $\widehat{\tau}$. In Table 2.2, $I = 5$, so $I(I+1)/4 = 5(5+1)/4 = 7.5$. Wilcoxon's T is $T = 8$ if computed from $Y_i - 1455.9999$, but it is $T = 7$ if computed from $Y_i - 1456.0001$, so $\widehat{\tau} = \$1456$. For all $I = 185$ pairs, the point estimate[8] of the increase in earnings is $\widehat{\tau} = \$1639$.

Although $\widehat{\tau}$ is one of the two well-known estimates that are called "Hodges–Lehmann (HL) estimates," Hodges and Lehmann [33] had actually proposed a general method for constructing estimates from tests of no treatment effect. The test statistic need not be the Wilcoxon statistic, and the parameter need not be an additive treatment effect. One equates a test statistic to its null expectation and solves the resulting equation for the point estimate. The method is applicable with other statistics. In a mildly dull way, if you take the sample mean of $Y_i - \tau_0$ as a test statistic, as in Sect. 2.3.2, then the mean of $Y_i - \tau_0$ has expectation zero when H_0 is true, so the estimating equation is $(1/I)\sum(Y_i - \widehat{\tau}) = 0$ with solution $\widehat{\tau} = \overline{Y}$; that is, the Hodges–Lehmann estimate derived from the mean is the mean of the Y_i. Almost, but not quite, the same thing happens with the randomization distribution of m-tests in Sect. 2.8. In Sect. 2.4.6 there are Hodges–Lehmann estimates of a multiplicative effect [6] and a Tobit effect.

2.4.4 The Average Treatment Effect

The average treatment effect is $\vartheta = (2I)^{-1}\sum_{i=1}^{I}\sum_{j=1}^{2}(r_{Tij} - r_{Cij})$ in a paired randomized experiment. Of course, ϑ cannot be calculated from the observed data, because the causal effect $r_{Tij} - r_{Cij}$ is not observed. If the causal effects, $r_{Tij} - r_{Cij}$, are well-defined—for instance, if there is no interference between units in Sect. 2.2.4—then ϑ exists and takes some numeric value. If the treatment effects were constant, $\tau = r_{Tij} - r_{Cij}$ for $i = 1, \ldots, I$ and $j = 1, 2$, then of course $\vartheta = \tau$, but ϑ exists as a number whether or not the treatment effects are constant.

[8] To compute the Hodges–Lehmann estimate in R, see footnote 7.

Can we estimate ϑ? Can we test the null hypothesis $H_0 : \vartheta = \vartheta_0$? Can we invert the hypothesis test to obtain a $1 - \alpha$ confidence set for ϑ?

In a paired randomized experiment, treatment assignment Z_{i1} is determined by the flip of a fair coin, with $Z_{i2} = 1 - Z_{i1}$, so $\Pr(Z_{ij} = 1|\mathscr{F}, \mathscr{Z}) = 1/2$ and $E\left(Z_{ij}|\mathscr{F}, \mathscr{Z}\right) = 1/2$. As a consequence, the treated-minus-control matched pair difference, $Y_i = (Z_{i1} - Z_{i2})(R_{i1} - R_{i2}) = Z_{i1}(r_{Ti1} - r_{Ci2}) + (1 - Z_{i1})(r_{Ti2} - r_{Ci1})$ has expectation $E(Y_i|\mathscr{F}, \mathscr{Z})$ equal to the average treatment effect in pair i, namely $(r_{Ti1} - r_{Ci1} + r_{Ti2} - r_{Ci2})/2$. Therefore, the mean difference, $\overline{Y}$ is unbiased for ϑ in a randomized experiment, $\vartheta = E\left(\overline{Y}|\mathscr{F}, \mathscr{Z}\right)$. That is the good news.

There is more good news, at least for short-tailed data in large paired randomized experiments, that is, as $I \to \infty$. If the mean $\overline{Y}$ is used as a test statistic in a paired randomized experiment, as in Table 2.5, and if the (r_{Tij}, r_{Cij}) are short-tailed like the Normal distribution, then a randomization test for a constant effect, $H_0 : \tau = \tau_0$, may safely be interpreted as a test of the same average effect, $H_0 : \vartheta = \tau_0$. The parallel $1 - \alpha$ confidence interval for a constant effect τ is also a $1 - \alpha$ confidence interval for the average effect ϑ. For one proof, see Baiocchi et al. [3, Proposition 2] (with their $d_{Tij} = 1$ and $d_{Cij} = 0$). See also the article by Peng Ding et al. [19].

The bad news, however, is that the mean, $\overline{Y}$, is not robust, as discussed in Sect. 2.3, and randomized experimentation does not change that. In particular, $\overline{Y}$ is unbiased for ϑ in a randomized experiment of size I pairs, but $\overline{Y}$ may be unstable even as $I \to \infty$. For instance, suppose that the r_{Cij} are $2I$ independent observations drawn from a Cauchy distribution, and suppose that the treatment effect is constant, $\tau = r_{Tij} - r_{Cij}$ for $i = 1, \ldots, I$ and $j = 1, 2$. Then $\vartheta = \tau$ for every I, and $\vartheta = E\left(\overline{Y}|\mathscr{F}, \mathscr{Z}\right)$ for every I, but as $I \to \infty$ the mean $\overline{Y}$ does not converge in probability to τ. However, the Hodges–Lehmann estimate $\widehat{\tau}$ of τ does converge in probability to τ. Indeed, under this model in a randomized experiment, before we condition on $\mathscr{F}$, the Y_i are I independent and identically distributed (iid) observations symmetrically distributed about τ, so inference about τ would be straightforward if robust methods were used in place of $\overline{Y}$.

There is more bad news. If we slightly change the model in the previous paragraph, then not only is $\overline{Y}$ unstable as $I \to \infty$, but the parameter ϑ itself is unstable. Suppose that $4I$ observations r_{Cij} and r'_{Cij} are independently drawn from a Cauchy distribution, and r_{Tij} is set to $\tau + r'_{Cij}$, discarding r'_{Cij} once r_{Tij} has been constructed. Then the unconditional distribution of the treated-minus-control pair difference Y_i in a randomized experiment is exactly the same as in the previous paragraph; in particular, before we condition on $\mathscr{F}$, the Y_i are iid and symmetric about τ. Once again, the Hodges–Lehmann estimate $\widehat{\tau}$ of τ does converge in probability to τ, but $\overline{Y}$ does not. Now, however, the average treatment effect, ϑ, changes as I changes, ϑ does not converge in probability to τ as $I \to \infty$, nor does $\overline{Y} - \vartheta$ converge to zero. Importantly, in this model, the I average treatment

effects, $(r_{Ti1} - r_{Ci1} + r_{Ti2} - r_{Ci2})/2$, in the I pairs vary from pair to pair but are symmetrically distributed about τ, and the Hodges–Lehmann estimate yields a consistent estimate of τ, but $\overline{Y}$ does not consistently estimate anything.

Indeed, in the model of the immediately previous paragraph in a randomized experiment, Wilcoxon's exact test of $H_0 : \tau = \tau_0$ falsely rejects a true null hypothesis at level α with exact probability at most α, and its associated $1 - \alpha$ exact confidence interval fails to cover τ with exact probability at most α. In this model, treatment effects that vary at random from pair to pair do not destabilize Wilcoxon's test or other analogous robust tests, but $\overline{Y}$ is not useful for inference. In parallel chapters, Lehmann [43, §3-§4] emphasizes that Wilcoxon's signed rank test yields the same valid inference in two very different sampling situations: randomized experimentation under Fisher's null hypothesis without iid sampling, or iid sampling of Y_i symmetric about zero without Fisher's null hypothesis. See also [44, Chapter 5].

In a paired randomized experiment, if the Y_i are asymmetrically distributed about their population median, then the treatment effect is not constant: some people are more strongly affected than others. Rather than obscuring this by taking an average, $\overline{Y}$, the investigator should take an interest in it, as discussed in subsequent sections and in Chap. 17.

2.4.5 Testing General Hypotheses About Treatment Effects

Any hypothesis that specifies the $2I$ treatment effects, $r_{Tij} - r_{Cij}$, may be tested using the same reasoning as in Sect. 2.4.1. Let τ_{0ij} be $2I$ specific numbers, and let $\boldsymbol{\theta}_0 = (\tau_{011}, \tau_{012}, \ldots, \tau_{0I2})^T$ collect them into a vector of dimension $2I$. Consider the hypothesis $H_0 : r_{Tij} = r_{Cij} + \tau_{0ij}$ for $i = 1, \ldots, I$ and $j = 1, 2$, or equivalently $H_0 : \mathbf{r}_T = \mathbf{r}_C + \boldsymbol{\theta}_0$. Any specific hypothesis about effects may be expressed in this way.

For instance, a basic hypothesis of this form entails interaction between the treatment and a covariate, or so-called effect-modification. In Table 2.1, one might entertain a hypothesis in which there are two values of τ_{0ij} depending upon whether the jth man in the ith pair has or does not have a high school degree. The hypothesis might assert that $\tau_{0ij} = \$1000$ if ij has a high school degree and $\tau_{0ij} = \$2000$ if ij does not have such a degree, so the program has a larger benefit for men without a degree.

The logic of Sect. 2.4.1 barely changes. If $H_0 : \mathbf{r}_T = \mathbf{r}_C + \boldsymbol{\theta}_0$ were true, the observed response $R_{ij} = Z_{ij}\, r_{Tij} + \left(1 - Z_{ij}\right) r_{Cij}$ from the jth person in pair i would be $R_{ij} = r_{Cij} + Z_{ij}\, \tau_{0ij}$, so that the observable data, $\left(R_{ij}, Z_{ij}\right)$, together with the hypothesis permit the computation of $r_{Cij} = R_{ij} - Z_{ij}\, \tau_{0ij}$, in parallel with Sect. 2.4.1. Indeed, using the fact that $Z_{i2} = 1 - Z_{i1}$, if $H_0 : \mathbf{r}_T = \mathbf{r}_C + \boldsymbol{\theta}_0$ were true, the treated-minus-control matched pair difference in responses is

$$\begin{aligned} Y_i &= (Z_{i1} - Z_{i2})\,(R_{i1} - R_{i2}) \\ &= (Z_{i1} - Z_{i2})\,\{(r_{Ci1} + \tau_{0i1}\,Z_{i1}) - (r_{Ci2} + \tau_{0i2}\,Z_{i2})\} \\ &= (Z_{i1} - Z_{i2})\,(r_{Ci1} - r_{Ci2}) + (\tau_{0i1}\,Z_{i1} + \tau_{0i2}\,Z_{i2}) \end{aligned}$$

so the adjusted difference

$$Y_i - (\tau_{0i1}\,Z_{i1} + \tau_{0i2}\,Z_{i2}) = (Z_{i1} - Z_{i2})\,(r_{Ci1} - r_{Ci2}) \tag{2.2}$$

is $\pm\,(r_{Ci1} - r_{Ci2})$ depending upon the random assignment of treatments $Z_{i1} - Z_{i2}$. The rest of the argument in Sect. 2.4.1 is as before, with a statistic, such as Wilcoxon's signed rank statistic, T, computed from the the adjusted differences (2.2), and compared to its randomization distribution formed from all 2^I assignments of signs to $(Z_{i1} - Z_{i2})\,(r_{Ci1} - r_{Ci2})$.

In abstract principle, a 95% confidence set for $\boldsymbol{\theta} = (r_{T11} - r_{C11}, \ldots, r_{TI2} - r_{CI2})^T$ could be constructed by testing every hypothesis $H_0 : \mathbf{r}_T = \mathbf{r}_C + \boldsymbol{\theta}_0$ and retaining for the confidence set the values of $\boldsymbol{\theta}_0$ that are not rejected by some 5% level test. In practice, such a confidence set would not be intelligible, for it would be a subset of $2I$-dimensional space. In the NSW data in Sect. 2.1, the confidence set would be a subset of $2I = 2 \times 185 = 370$ dimensional space and would be beyond human comprehension. In brief, it is straightforward to make valid statistical inferences that are so complex, so faithful to the minute detail of reality, that they are unintelligible and of no practical use whatsoever. The 1-dimensional confidence interval for the single (scalar) constant effect τ in Sect. 2.4.2 provides insight into the $2I$-dimensional confidence set for $\boldsymbol{\theta}$ because the constant model $\mathbf{r}_T = \mathbf{r}_C + \mathbf{1}\tau$ defines a line in $2I$-dimensional space as τ varies over the real numbers. If all the tests were performed using the same test statistic, say Wilcoxon's signed rank statistic T, then τ_0 is excluded from the 1-dimensional confidence interval for an additive effect if and only if $\boldsymbol{\theta}_0 = (\tau_0, \tau_0, \ldots, \tau_0)^T$ is excluded from the $2I$-dimensional confidence set for $\boldsymbol{\theta}$; after all, it is the same hypothesis tested by the same test. In this sense, a 1-dimensional model for the $2I$ dimensional effect, such as the constant effect model, may be understood as an attempt to glean insight into the $2I$ dimensional confidence set for $\boldsymbol{\theta}$ while recognizing that any 1-dimensional model, indeed any intelligible model, is to some degree an oversimplification. Understanding of $\boldsymbol{\theta}$ is often aided by contrasting several intelligible models, rather than discarding them. Two other 1-dimensional effects are considered in Sect. 2.4.6. Arguably, the joint consideration of three 1-dimensional models for the $2I$-dimensional parameter $\boldsymbol{\theta}$ provides more humanly accessible insight into $\boldsymbol{\theta}$ than would a $2I$-dimensional confidence set.

An alternative approach to understanding the $2I$-dimensional confidence set for $\boldsymbol{\theta}$ uses "attributable effects." In typical practice, an attributable effect provides a 1-dimensional summary statement about the $2I$-dimensional $\boldsymbol{\theta}$, but this summary does not determine $\boldsymbol{\theta}$ uniquely. Attributable effects are discussed in Sect. 2.5.

2.4.6 Multiplicative Effects; Tobit Effects

A Constant Multiplicative Effect β

As noted in Sect. 2.4.1, the hypothesis of a positive, constant additive effect, $H_0 : r_{Tij} = r_{Cij} + \tau$ for $i = 1, \ldots, I$, $j = 1, 2$, has the disagreeable property that it implies that a man with zero earnings under treatment would have negative earnings under control. In this section, two other 1-dimensional families of hypotheses for the $2I$-dimensional effect $\boldsymbol{\theta} = (r_{T11} - r_{C11}, \ldots, r_{TI2} - r_{CI2})^T$ are considered; they do not produce negative earnings.

The first hypothesis asserts that the effect is a constant multiplier rather than a constant addition, $H_0 : r_{Tij} = \beta_0 r_{Cij}$ for $i = 1, \ldots, I$, $j = 1, 2$ with $\beta_0 \geq 0$. This hypothesis avoids the implication of negative earnings under one of the treatments. The multiplicative effect hypothesis is a special case of the hypothesis in Sect. 2.4.5, namely the hypothesis $H_0 : \mathbf{r}_T = \mathbf{r}_C + \boldsymbol{\theta}_0$ with $\tau_{0ij} = r_{Tij} - r_{Cij} = \beta_0 r_{Cij} - r_{Cij} = (\beta_0 - 1)\, r_{Cij}$.

For an additive effect, the treatment effect is removed by subtraction, but for a multiplicative effect, the effect is removed by division [6]. If $H_0 : r_{Tij} = \beta_0 r_{Cij}$ were true for all ij, then $R_{ij}/\beta_0^{Z_{ij}} = r_{Cij}$; that is, if $Z_{ij} = 0$, then $R_{ij} = r_{Cij}$ and $\beta_0^{Z_{ij}} = 1$, while if $Z_{ij} = 1$, then $R_{ij} = r_{Tij}$, $\beta_0^{Z_{ij}} = \beta_0$ and $R_{ij}/\beta_0^{Z_{ij}} = r_{Tij}/\beta_0 = r_{Cij}$. The hypothesis $H_0 : r_{Tij} = \beta_0 r_{Cij}$ is tested by applying Wilcoxon's signed rank statistic to the I adjusted, treated-minus-control matched pair differences, $(Z_{i1} - Z_{i2})\left(R_{i1}/\beta_0^{Z_{i1}} - R_{i2}/\beta_0^{Z_{i2}}\right)$, which equal $(Z_{i1} - Z_{i2})\,(r_{Ci1} - r_{Ci2})$ if the hypothesis is true. The set of values of β_0 not rejected in a 5%-level test is a 95% confidence set for a constant multiplicative effect β.

In the NSW experiment with $I = 185$ pairs, the two-sided, 95% confidence interval for β is [1.08, 1.95], or between an 8% increase and a 95% increase. The Hodges–Lehmann point estimate is $\widehat{\beta} = 1.45$.

Although a multiplicative effect avoids the implication of negative earnings, it has peculiarities of its own. If the effect is multiplicative, then a man who has $r_{Cij} = 0$ earnings under control also has $r_{Tij} = \beta\, r_{Cij} = 0$ earnings under treatment. Many controls did have zero earnings in 1978; see Fig. 2.1. Among its several goals, the NSW program intended to bring into employment men who would otherwise be out of work, and a multiplicative effect says this cannot happen. This leads to the consideration of a "Tobit effect."

Tobit Effects

Two men with the same positive earnings, r_{Cij}, under control in 1978 have something in common: the labor market attached the same dollar value to their labor. Two men with zero earnings under control in 1978 have less in common: the labor market declined to attach a positive value to the labor they might supply. One might entertain the thought that two unemployed men are unequally unemployed:

perhaps the market attaches negative but unequal values to the labor they might supply. Expressed in operational terms, the subsidy that an employer would require before employing the two unemployed men might be different because one man could supply labor of greater net value to the employer than the other. In a simple economic model, workers are paid the values of the marginal products of their labor, and a worker is unemployed if the marginal product of his labor is negative. (We have all known people like that.) "Tobit effects," named for James Tobin [79], attempt to express this idea.

Suppose the NSW program raises the marginal value of every worker's labor by the same constant, τ_0, but the worker is employed, with positive earnings, only if the marginal value of the labor is positive. In this case, if the jth man in the ith pair would have earnings of r_{Tij} in the NSW program, and if $r_{Tij} \geq \tau_0$, then $r_{Cij} = r_{Tij} - \tau_0$, but if $r_{Tij} < \tau_0$ then $r_{Cij} = 0$; that is, in general, $r_{Cij} = \max\left(r_{Tij} - \tau_0, 0\right)$. This hypothesis is another special case of the general hypothesis in Sect. 2.4.5.

If the hypothesis, $H_0 : r_{Cij} = \max\left(r_{Tij} - \tau_0, 0\right)$ for $i = 1, \ldots, I$, $j = 1, 2$, were true, then $\max\left(R_{ij} - \tau_0 Z_{ij},\ 0\right) = r_{Cij}$. The hypothesis $r_{Cij} = \max\left(r_{Tij} - \tau_0, 0\right)$ may be tested by applying Wilcoxon's signed rank test to the I adjusted treated-minus-control differences,

$$(Z_{i1} - Z_{i2})\left\{\max\left(R_{i1} - \tau_0 Z_{i1},\ 0\right) - \max\left(R_{i2} - \tau_0 Z_{i2},\ 0\right)\right\},$$

which equal $(Z_{i1} - Z_{i2})\,(r_{Ci1} - r_{Ci2})$ if the hypothesis is true. The test is the same as before; see Sect. 2.4.5.

For the Tobit effect $r_{Cij} = \max\left(r_{Tij} - \tau, 0\right)$, the 95% confidence interval for τ is $\left[\$458,\ \$3955\right]$, and the Hodges–Lehmann point estimate is $\widehat{\tau} = \$2114$.

Comparing the Three Effects: Additive, Multiplicative, Tobit

Figure 2.2 compares an additive effect, $r_{Cij} = r_{Tij} - \tau$, a multiplicative effect, $r_{Cij} = r_{Tij}/\beta$, and a Tobit effect, $r_{Cij} = \max\left(r_{Tij} - \tau, 0\right)$. In each of the three panels of Fig. 2.2, the 1978 earnings of the 185 treated men ($Z_{ij} = 1$) and the 185 control men ($Z_{ij} = 0$) are displayed after removing the estimated treatment effect from the earnings of the treated men. In each case, the Hodges–Lehmann estimate is used to estimate the treatment effect. For an additive effect, $R_{ij} - \widehat{\tau} Z_{ij}$ is plotted; for a multiplicative effect, $R_{ij}/\widehat{\beta}^{Z_{ij}}$ is plotted; for a Tobit effect, $\max\left(R_{ij} - \widehat{\tau}\, Z_{ij},\ 0\right)$ is plotted. If the number I of pairs were large, then the pair of treated-control boxplots would look the same with an effect of the correct form. The boxplots for an additive effect suggest the effect is not, in fact, additive, in part because the boxplots look quite different, and in part because negative earnings are produced. The boxplots for the multiplicative effect look better, although the lower quartile for the treated group is positive while for the control group it is zero. The boxplots for the Tobit effect are better than for the additive effect, and the lower quartile in both groups is now zero. Although the Tobit effect is, arguably, the best of the three, it

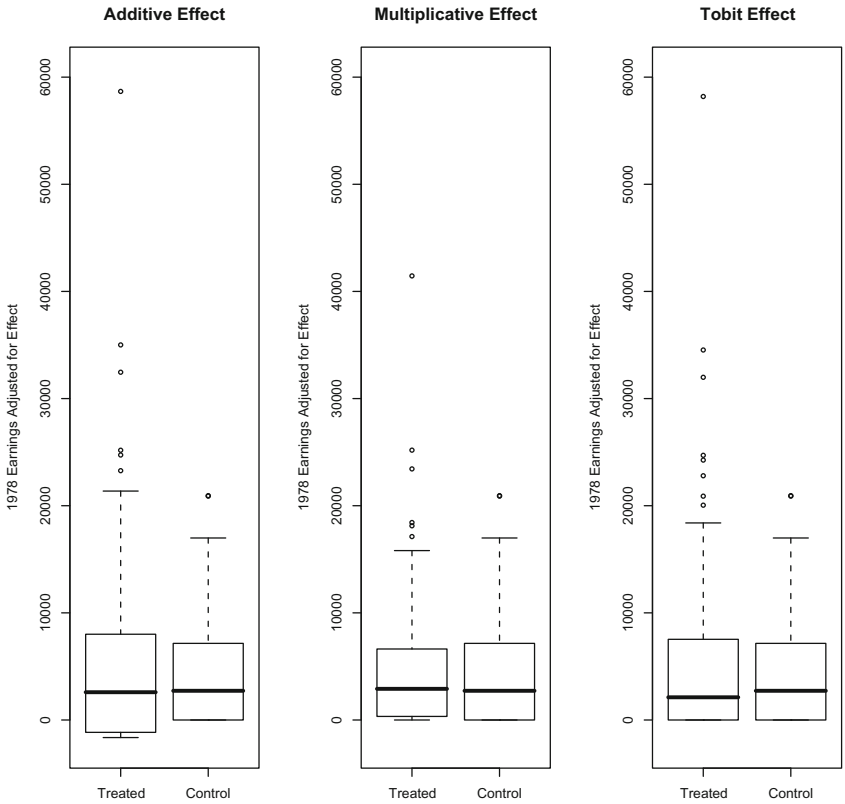

Fig. 2.2 Earnings in 1978 after adjustment to remove three possible treatment effects, an additive effect of \$1639, a multiplicative effect of 1.45, and a Tobit effect of \$2114. If the correct effect were removed, the two boxplots would look similar, differing only by chance

misses the long right tail in the treated group: a handful of treated men earned much more than is accounted for by the Tobit effect. This was clearly visible in Fig. 2.1 as well. More will be said about the tail in Sect. 2.5.

2.5 Attributable Effects

2.5.1 Why Use Attributable Effects?

As seen in Sect. 2.4.5, it is straightforward to test a hypothesis that specifies all $2I$ treatment effects, $r_{Tij} - r_{Cij}$, say the hypothesis $H_0 : r_{Tij} = r_{Cij} + \tau_{0ij}$ for $i = 1, \ldots, I$ and $j = 1, 2$, or equivalently the hypothesis $H_0 : \mathbf{r}_T = \mathbf{r}_C + \boldsymbol{\theta}_0$ with $\boldsymbol{\theta}_0 = (\tau_{011}, \tau_{012}, \ldots, \tau_{0I2})^T$. We were barred from using such a test to produce a

confidence set for $\boldsymbol{\theta} = (r_{T11} - r_{C11}, \ldots, r_{TI2} - r_{CI2})^T$ in part by our inability to make practical sense of a confidence set that is a subset of $2I$-dimensional space. In Sect. 2.4.6, some insight into $\boldsymbol{\theta}$ was gleaned by contrasting three 1-dimensional effects, that is, three paths or curves through the $2I$-dimensional space of possible effects. By contrast, "attributable effects" make summary statements about $\boldsymbol{\theta}$ without restricting the form of $\boldsymbol{\theta}$ in any way. Attributable effects are a general concept [57, 58] but, in the current section, only the case of matched pairs is considered using Wilcoxon's signed rank statistic [61] and an extension of it by W. Robert Stephenson [64, 77]. The discussion is simplified by assuming that there are no ties of any kind and discussing ties separately at the end.

2.5.2 Aligned Responses: Shifting Attention from Pairs to Individuals

If no simplifying form is imposed upon $\boldsymbol{\theta} = (r_{T11} - r_{C11}, \ldots, r_{TI2} - r_{CI2})$, then it is easiest to understand a treatment effect on one man, $r_{Tij} - r_{Cij}$, rather than on the matched pair difference between two men. In simple cases, such as an additive effect, the difference of two matched men has a simple relationship with the separate responses of the two individual men (Sect. 2.4.1), but in general this is not so. At the same time, the men were carefully matched for important characteristics, such as education, age, and pretreatment earnings, and it seems wise to retain the comparison of two similar men. These two considerations lead to "aligned responses," introduced by Hodges and Lehmann [32] in an extension of Wilcoxon's signed rank statistic. With matched pairs, the aligned response for the jth man in the ith pair is the difference between his observed response, R_{ij}, and the average response in his pair, $\overline{R}_i = (R_{i1} + R_{i2})/2$, so the aligned response is $R_{ij} - \overline{R}_i$. Talk about aligned responses is, in a sense, talk about one man while taking account of the pairing. Of course, the treated-minus-control pair difference of aligned responses $R_{ij} - \overline{R}_i$ equals the treated-minus-control difference in the responses R_{ij} themselves because $\overline{R}_i$ appears twice and cancels. If Wilcoxon's signed rank statistic were computed from the aligned responses, it would equal Wilcoxon's signed rank statistic computed from the responses themselves. So far, the switch to aligned responses is a switch in the way we talk rather than a switch in what we calculate.[9]

[9] At this point, we part company with Hodges and Lehmann [32]. They were interested in creating a generalization of Wilcoxon's signed rank statistic so that the matched sets were no longer pairs and might, for example, include several controls. Their statistic, the aligned rank statistic, ranks the aligned responses (not the absolute differences) and sums the ranks of treated subjects. For matched pairs, they would rank the $R_{ij} - \overline{R}_i$ from 1 to $2I$ and sum the I ranks for the I treated subjects. Hodges and Lehmann [32] show that, for matched pairs, Wilcoxon's signed rank statistic and the aligned rank statistic produce virtually the same inferences. Can you see why? In pairs, how are $R_{i1} - \overline{R}_i$ and $R_{i2} - \overline{R}_i$ related? Think about the pair with largest $R_{ij} - \overline{R}_i$, the pair that

If the null hypothesis of no treatment effect, $H_0 : r_{Tij} = r_{Cij}$ for $i = 1, \ldots, I$ and $j = 1, 2$, were true, then the pair of aligned responses in pair i would be

$$\left(r_{Ci1} - \frac{r_{Ci1} + r_{Ci2}}{2},\ r_{Ci2} - \frac{r_{Ci1} + r_{Ci2}}{2}\right) \tag{2.3}$$

no matter how treatments were assigned. In contrast, if the treatment has an effect, then the pair of aligned responses in pair i would be

$$\left(r_{Ti1} - \frac{r_{Ti1} + r_{Ci2}}{2},\ r_{Ci2} - \frac{r_{Ti1} + r_{Ci2}}{2}\right) \tag{2.4}$$

if the first man were the treated man, $Z_{i1} = 1$, $Z_{i2} = 0$, and the aligned responses would be

$$\left(r_{Ci1} - \frac{r_{Ci1} + r_{Ti2}}{2},\ r_{Ti2} - \frac{r_{Ci1} + r_{Ti2}}{2}\right) \tag{2.5}$$

if the second man were the treated man, $Z_{i1} = 0$, $Z_{i2} = 1$, so in either case the pair of aligned responses is

$$\begin{aligned}\left(R_{i1} - \overline{R}_i,\ R_{i2} - \overline{R}_i\right) &= Z_{i1}\left(r_{Ti1} - \frac{r_{Ti1} + r_{Ci2}}{2},\ r_{Ci2} - \frac{r_{Ti1} + r_{Ci2}}{2}\right) \\ &\quad + Z_{i2}\left(r_{Ci1} - \frac{r_{Ci1} + r_{Ti2}}{2},\ r_{Ti2} - \frac{r_{Ci1} + r_{Ti2}}{2}\right).\end{aligned} \tag{2.6}$$

For instance, it might happen that $r_{Ti1} - (r_{Ti1} + r_{Ci2})/2 = \$1000$ in (2.4) and $r_{Ti2} - (r_{Ci1} + r_{Ti2})/2 = \$5000$ in (2.5), so the effect in pair i is positive and substantial but far from constant; rather it depends upon which man is treated. Can Wilcoxon's statistic (and Stephenson's generalization) help to draw inferences about effects that vary in size from person to person?

2.5.3 *Thinking About Heterogeneous Effects by Focusing on Small Groups of Pairs*

To think about effects that are not constant, that are heterogeneous, it is natural to look at several pairs at once. If the effect varies, variations in effect have to part of the story. Let $\mathscr{I} \subseteq \{1, \ldots, I\}$ be a subset of m pairs of the I pairs. Wilcoxon's statistic looks at pairs $m = 2$ at a time, so $\mathscr{I} = \left\{i, i'\right\}$ for $i \neq i'$, $1 \leq i < i' \leq I$.

contains the aligned response that gets the largest aligned rank. How is that pair related to the pair that has the largest $|R_{i1} - R_{i2}|$, the one that has the largest rank in Wilcoxon's statistic?

Stephenson [77] suggested using other values of m, $1 \leq m \leq I$. For the moment, focus on one subset $\mathscr{I}$, perhaps $\mathscr{I} = \{1, 2\}$; later, the statistic will average over all subsets.

Among the m pairs in $\mathscr{I}$, there are $2m$ men, or $2m = 4$ men for Wilcoxon's statistic. Which one man among these $2m$ men seems to have the most exceptionally positive earnings, in the sense that his observed aligned response, $R_{ij} - \overline{R}_i$, is the largest, the most positive, of the $2m$ aligned responses? That one man seems to be an interesting man, deserving special attention, because he entered or reentered the workforce with high earnings in 1978 relative to someone else of similar age, education and so forth. This man is the most interesting of $2m$ men. Now that we have focused on one interesting man, it is time to ask: was he a treated man? We ought to tally one success for the NSW treatment if this man—the man with the highest aligned earnings among the $2m$ men in $\mathscr{I}$—is a treated man. So write $H_{\mathscr{I}} = 1$ if indeed he is a treated man, and $H_{\mathscr{I}} = 0$ if he is a control.

Even if the treatment had no effect, so that the aligned responses were always (2.3), the random assignment of treatments would grant a success, $H_{\mathscr{I}} = 1$, to the treatment with probability $\frac{1}{2}$. One of the two values in the pair (2.3) is positive, and even if the treatment is without effect, randomization picks the positive entry for treatment half the time. That is true for each pair, so it is true for the pair $i \in \mathscr{I}$ with the largest $r_{Cij} - (r_{Ci1} + r_{Ci2})/2$ in (2.3). Evidently, some work needs to be done to distinguish apparent successes that are merely due to luck of the random assignment and successes that were actually caused by the treatment.

To distinguish successes caused by treatment from successes that happen by luck, a new quantity, $H_{C\mathscr{I}}$, is introduced. Imagine, for a brief moment, that we could observe the response to control, r_{Cij}, for all $2I$ subjects, so that, in particular, we could calculate the pair of aligned control responses in (2.3). In point of fact, we observe only half of the r_{Cij} because half of the men receive treatment, so we observe (2.4) or (2.5) not (2.3). If we could calculate (2.3), we could calculate a quantity $H_{C\mathscr{I}}$ which is similar to $H_{\mathscr{I}}$, but is computed from the aligned control responses in (2.3). Specifically, with (2.3) in hand, we could find the man with the largest, most positive of the $2m$ aligned responses to control (2.3) for the m pairs $i \in \mathscr{I}$ and score $H_{C\mathscr{I}} = 1$ if that man was a treated man and $H_{C\mathscr{I}} = 0$ if that man was a control. Like $H_{\mathscr{I}}$, the quantity $H_{C\mathscr{I}}$ depends on the random assignment of treatments, $\mathbf{Z}$, so $H_{\mathscr{I}}$ and $H_{C\mathscr{I}}$ are random variables; however, $H_{\mathscr{I}}$ is observed but $H_{C\mathscr{I}}$ is not.

If $H_{\mathscr{I}} - H_{C\mathscr{I}} = 1 - 0 = 1$, then in the set $\mathscr{I}$ an effect of the treatment caused a success to be tallied that would not have been tallied if all $2m$ men had received the control. If $H_{\mathscr{I}} - H_{C\mathscr{I}} = 0 - 1 = -1$, then in the set $\mathscr{I}$ an effect of the treatment prevented a success from being tallied that would have been tallied if all $2m$ men had received the control. If $H_{\mathscr{I}} - H_{C\mathscr{I}} = 0$, then in the set $\mathscr{I}$ the effects of the treatment did not change whether a success was tallied. One might say, with some imprecision, that $H_{\mathscr{I}}$ gives credit for successes that occur by chance, but $H_{\mathscr{I}} - H_{C\mathscr{I}}$ does not. We cannot calculate quantities like $H_{\mathscr{I}} - H_{C\mathscr{I}}$ because we see $H_{\mathscr{I}}$ but not $H_{C\mathscr{I}}$; however, if we could calculate $H_{\mathscr{I}} - H_{C\mathscr{I}}$, we would know

whether a success or failure was attributable to effects caused by the treatment. That we never see $H_{C\mathscr{I}}$ turns out to be less of a problem than one might suppose.

It seems arbitrary to focus on just the $2m$ men in $\mathscr{I}$ when there are data on $2I$ men. For Wilcoxon's statistic, with $m = 2$, this would mean focusing on $2m = 4$ men when there are $2I = 370$ men to consider. To drive out the arbitrary element of what otherwise seems to be a promising idea, it is natural to consider every subset $\mathscr{I}$ of m distinct pairs, to compute $H_{\mathscr{I}}$ for each one, and to tally up the total number of successes for the treatment in all of these comparisons. So let $\mathscr{K}$ be the collection of all subsets $\mathscr{I}$ of m distinct pairs. For Wilcoxon's statistic, with $m = 2$, the collection is $\mathscr{K}$ contains $\{1, 2\}, \{1, 3\}, \{2, 3\}, \ldots, \{I - 1, I\}$. In general, $\mathscr{K}$ contains $|\mathscr{K}| = I!/\{m!\,(I - m)!\}$ subsets, or $17,020$ subsets of two pairs for the Wilcoxon statistic in the NSW experiment. The statistic $\widetilde{T} = \sum_{\mathscr{I}\in\mathscr{K}} H_{\mathscr{I}}$ tallies up the successes for the treatment in all possible comparisons of m distinct pairs.[10] In a parallel, the unobservable quantity $\widetilde{T}_C = \sum_{\mathscr{I}\in\mathscr{K}} H_{C\mathscr{I}}$ tallies the success for the treatment that would have occurred had all $2I$ men received the control. Finally, $A = \widetilde{T} - \widetilde{T}_C = \sum_{\mathscr{I}\in\mathscr{K}} H_{\mathscr{I}} - H_{C\mathscr{I}}$ tallies the number of successes attributable to effects of the treatment, that is, the net increase in successes due to effects of the treatment. If the treatment has no effect, then $H_{\mathscr{I}} = H_{C\mathscr{I}}$ for every $\mathscr{I} \in \mathscr{K}$ and $A = 0$. In general, A is an integer between $-|\mathscr{K}|$ and $|\mathscr{K}|$. It is often convenient to think of the proportions, $\widetilde{T}/|\mathscr{K}|$, $\widetilde{T}_C/|\mathscr{K}|$, and $A/|\mathscr{K}|$ rather than the counts.

2.5.4 Computations; Null Distribution

Computing $\widetilde{T}$ sounds as if it would be an enormous amount of work: even for $m = 2$, there are $17,020$ terms $H_{\mathscr{I}}$ for the $I = 185$ pairs in the NSW experiment. Actually, it is easy. Stephenson [77] wrote $\widetilde{T}$ in the form of a signed rank statistic, similar to the form for Wilcoxon's T in Sect. 2.3.3, specifically $\widetilde{T} = \sum_{i=1}^{I} \operatorname{sgn}\{(Z_{i1} - Z_{i2})\,(r_{Ci1} - r_{Ci2})\} \cdot \widetilde{q}_i$, where $\operatorname{sgn}(a) = 1$ if $a > 0$, $\operatorname{sgn}(a) = 0$ if $a \le 0$, and $\widetilde{q}_i$ is a new quantity that will be defined in just a moment. With pairs, it is always true that $R_{i1} - \overline{R}_i = -R_{i2} + \overline{R}_i$, so $\left|R_{i1} - \overline{R}_i\right| = \left|R_{i2} - \overline{R}_i\right|$. Rank the pairs from 1 to I based on the value of $\left|R_{i1} - \overline{R}_i\right|$; therefore, pair i gets rank q_i where this rank q_i is the same as the rank q_i in Sect. 2.3.3 of $|Y_i|$ that appears in Wilcoxon's signed rank statistic. The pairs with ranks $q_i = 1$, $q_i = 2, \ldots,$ $q_i = m - 1$, never contain the largest observed aligned response $R_{ij} - \overline{R}_i$ in a subset of m pairs, so set $\widetilde{q}_i = 0$ for these pairs. The pair with rank $q_i = m$ contains

[10] A statistic built in this way is known as a U-statistic, a type of statistic invented by Wassily Hoeffding [34], and for $m = 2$, the statistic $\widetilde{T}$ is the U-statistic that most closely resembles Wilcoxon's signed rank statistic; see [43, Chapter 4] or [53, §3.5]. Wilcoxon's statistic T does not quite equal $\widetilde{T}$, but the behavior of the two statistics is virtually identical when I is large. If Wilcoxon had ranked the I pairs from 0 to $I - 1$, he would have produced $\widetilde{T}$ rather than T [53, §3.5]. If the collection $\mathscr{K}$ were expanded to include the I self-pairs, $(1, 1), (2, 2), \ldots, (I, I)$, then $\widetilde{T}$ would equal T [43, Chapter 4].

the largest observed aligned response $R_{ij} - \overline{R}_i$ in exactly one subset of m pairs, so set $\widetilde{q}_i = 1$ for this pair. The pair with rank $q_i = m + 1$ contains the largest observed aligned response $R_{ij} - \overline{R}_i$ in exactly m subsets $\mathscr{I}$ of m pairs, so set $\widetilde{q}_i = m$ for this pair. In general, the pair with rank q_i contains the largest aligned response in exactly $\widetilde{q}_i$ pairs, where

$$\widetilde{q}_i = \binom{q_i - 1}{m - 1} \text{ for } q_i \geq m, \quad \widetilde{q}_i = 0 \text{ for } q_i < m. \tag{2.7}$$

With $\widetilde{q}_i$ so defined,

$$\begin{aligned} \widetilde{T} &= \sum_{\mathscr{I} \in \mathscr{K}} H_{\mathscr{I}} \\ &= \sum_{i=1}^{I} \operatorname{sgn}\{(Z_{i1} - Z_{i2})(r_{Ci1} - r_{Ci2})\} \cdot \widetilde{q}_i \,. \end{aligned}$$

With $m = 2$ for Wilcoxon's statistic,

$$\binom{q_i - 1}{m - 1} = \binom{q_i - 1}{1} = q_i - 1, \tag{2.8}$$

so T and $\widetilde{T}$ are very similar, except the ranks in $\widetilde{T}$ are $0, 1, \ldots, I - 1$ rather than $1, 2, \ldots, I$; see footnote 10 and [53, §3.5]. With $m = 5$, the ranks $\widetilde{q}_i$ are 0, 0, 0, 0, 1, 5, 15, 35, 70, 126, ...

If the null hypothesis of no treatment effect, $H_0 : r_{Tij} = r_{Cij}$ for $i = 1, \ldots, I$ and $j = 1, 2$, were true, then the exact null distribution of $\widetilde{T}$ could be determined by direct enumeration, as was done for Wilcoxon's T in Tables 2.7 and 2.8. If H_0 is true, then $\widetilde{T}$ has the distribution of the sum of I independent random variables taking the values 0 and $\widetilde{q}_i$ each with probability $\frac{1}{2}$, for $i = 1, \ldots, I$. Also, the central limit theorem would provide a Normal approximation to the null distribution of $\widetilde{T}$ based on its null expectation and variance.

This null distribution of $\widetilde{T}$ shares with Wilcoxon's statistic the curious property noted in Sect. 2.3.3, namely that, in the absence of ties, the null distribution may be written down before the experiment is conducted, without seeing any of the responses. This turns out to be convenient. In particular, even though we cannot observe the $2I$ potential responses to control in (2.3), we do know that, by definition, they satisfy the null hypothesis of no treatment effect. In light of this, even though we cannot observe any of the $H_{C\mathscr{I}}$, we do know the distribution of

$$\widetilde{T}_C = \sum_{\mathscr{I} \in \mathscr{K}} H_{C\mathscr{I}}. \tag{2.9}$$

The distribution of the unobservable $\widetilde{T}_C$ is the known null distribution of $\widetilde{T}$. That is, even when the treatment has an effect, in the absence of ties, the distribution of $\widetilde{T}_C$ is available by direct enumeration with the ranks, $\widetilde{q}_i$; moreover,

$$E\left(\widetilde{T}_C \middle| \mathscr{F}, \mathscr{Z}\right) = (1/2) \sum_{i=1}^{I} \widetilde{q}_i, \tag{2.10}$$

$$\operatorname{var}\left(\widetilde{T}_C \middle| \mathscr{F}, \mathscr{Z}\right) = (1/4) \sum_{i=1}^{I} \widetilde{q}_i^2 \tag{2.11}$$

and the distribution of $\left\{\widetilde{T}_C - E\left(\widetilde{T}_C \middle| \mathscr{F}, \mathscr{Z}\right)\right\} / \sqrt{\operatorname{var}\left(\widetilde{T}_C \middle| \mathscr{F}, \mathscr{Z}\right)}$ converges in distribution to the standard Normal distribution as $I \to \infty$ with m fixed. Using this known distribution of $\widetilde{T}_C$, find the smallest value t_α such that $\Pr\left(\widetilde{T}_C \leq t_\alpha \middle| \mathscr{F}, \mathscr{Z}\right) \geq 1 - \alpha$. For large I, the critical value t_α is approximately

$$E\left(\widetilde{T}_C \middle| \mathscr{F}, \mathscr{Z}\right) + \Phi^{-1}(1-\alpha) \sqrt{\operatorname{var}\left(\widetilde{T}_C \middle| \mathscr{F}, \mathscr{Z}\right)}, \tag{2.12}$$

where $\Phi^{-1}(\cdot)$ is the standard Normal quantile function, or

$$E\left(\widetilde{T}_C \middle| \mathscr{F}, \mathscr{Z}\right) + 1.65 \sqrt{\operatorname{var}\left(\widetilde{T}_C \middle| \mathscr{F}, \mathscr{Z}\right)} \tag{2.13}$$

for $\alpha = 0.05$.

Our interest is in the number $A = \widetilde{T} - \widetilde{T}_C$ of successes attributable to effects of the treatment, where $\widetilde{T}$ can be calculated from the data and $\widetilde{T}_C$ is not observed. It is easy to see that $\Pr\left(A \geq \widetilde{T} - t_\alpha \middle| \mathscr{F}, \mathscr{Z}\right) \geq 1 - \alpha$, where $\widetilde{T} - t_\alpha$ is something that can be calculated, so $A \geq \widetilde{T} - t_\alpha$ may be asserted with $1 - \alpha$ confidence.[11] In terms of proportions of successes rather than counts, the confidence statement is $A / |\mathscr{K}| \geq \left(\widetilde{T} - t_\alpha\right) / |\mathscr{K}|$.

Ties

Ties are a minor inconvenience. Among the $I = 185$ pairs in the NSW experiment, there are 14 ties among the $|Y_i|$, all with $|Y_i| = 0$. Briefly, ties make the procedure conservative, so $A \geq \widetilde{T} - t_\alpha$ occurs with probability at least $1 - \alpha$ for large I.

[11] To see this, observe that: $\Pr\left(A \geq \widetilde{T} - t_\alpha \middle| \mathscr{F}, \mathscr{Z}\right) = \Pr\left(\widetilde{T} - \widetilde{T}_C \geq \widetilde{T} - t_\alpha \middle| \mathscr{F}, \mathscr{Z}\right) = \Pr\left(\widetilde{T}_C \leq t_\alpha \middle| \mathscr{F}, \mathscr{Z}\right) \geq 1 - \alpha$ by the definition of t_α. The assertion $A \geq \widetilde{T} - t_\alpha$ is a confidence interval for an unobservable random variable. Confidence sets for random variables are less familiar than confidence sets for parameters, but they are hardly new; see, for instance, Weiss [82].

Specifics follow. Although the distribution of $\widetilde{T}_C$ is known in the absence of ties, it is not known if ties occur, because the pattern of ties in the unobservable (2.3) affects the distribution. For the observable $\widetilde{T}$, it is natural to use average ranks, averaging the $\widetilde{q}_i$ for pairs with tied $|Y_i|$, and setting $\operatorname{sgn}(a) = 1$ for $a > 0$, $\operatorname{sgn}(a) = \frac{1}{2}$ for $a = 0$, $\operatorname{sgn}(a) = 0$ for $a < 0$. Speaking informally, a tie for the highest $R_{i1} - \overline{R}_i$ for $i \in \mathscr{I}$ means one success is shared as a fractional success equally among the tied values. The same thing cannot be done for $\widetilde{T}_C$ because the ties in (2.3) are not observed. Suppose that the unobserved $\widetilde{T}_C$ allowed for ties in the manner proposed for the observed $\widetilde{T}$, but the distribution used to determine t_α ignored ties, using instead the ranks $\widetilde{q}_i$ in (2.7) for $q_i = 1, 2, \ldots, I$; then the expectation is correct, $E\left(\widetilde{T}_C \middle| \mathscr{F}, \mathscr{Z}\right) = (1/2)\sum_{i=1}^{I} \widetilde{q}_i$ or $E\left(\widetilde{T}_C / |\mathscr{K}| \middle| \mathscr{F}, \mathscr{Z}\right) = \frac{1}{2}$, the variance is too large, $\operatorname{var}\left(\widetilde{T}_C \middle| \mathscr{F}, \mathscr{Z}\right) \leq (1/4)\sum_{i=1}^{I} \widetilde{q}_i^2$, the approximate $t_\alpha = E\left(\widetilde{T}_C \middle| \mathscr{F}, \mathscr{Z}\right) + \Phi^{-1}(1-\alpha)\sqrt{\operatorname{var}\left(\widetilde{T}_C \middle| \mathscr{F}, \mathscr{Z}\right)}$ is too large for large I and $\alpha < \frac{1}{2}$, so $\Pr\left(\widetilde{T}_C \leq t_\alpha \middle| \mathscr{F}, \mathscr{Z}\right) \geq 1 - \alpha$, and $A \geq \widetilde{T} - t_\alpha$ may be asserted with at least $1 - \alpha$ confidence. For $m = 2$ in the NSW experiment, the ratio of the tied to untied variances for $\widetilde{T}$ is 0.9996 for $m = 2$, and is closer to 1 for larger m.

Results for the NSW Experiment: Big Effects, Now and Then

Using the Wilcoxon signed rank statistic means looking at pairs $m = 2$ at a time. For the $I = 185$ pairs, there are 17,020 subsets of two pairs. In 60.8% of these 17,020 subsets, the man with the highest aligned earnings $R_{i1} - \overline{R}_i$ was a treated man, whereas 50% is expected by chance, so the point estimate is a 10.8% increase attributable to effects of the treatment, but we are 95% confident of only a 3.8% increase.

More interesting is what happens in groups of $m = 20$ pairs or $2m = 40$ men. Here, there are 3.1×10^{26} comparisons $\mathscr{I}$ in $\mathscr{K}$. In 85.7% of these comparisons, the man with the highest aligned earnings $R_{i1} - \overline{R}_i$ was a treated man, whereas 50% were expected by chance, so the point estimate is a 35.7% increase attributable to effects of the treatment, but we are 95% confident of only a 15.8% increase. When you look at 40 matched men and you pick the man with the most impressive increase in aligned earnings, that man is typically a treated man: it happened 85.7% of the time. When there are big gains in earnings, mostly they occur in the treated group. As Fig. 2.1 suggests, the treatment has a big effect on a subset of men, with perhaps little effect on many others. Indeed, the possibility of detecting large but rare effects is the motivation for considering $m > 2$.

Uncommon But Dramatic Responses to Treatment

If the effect of a treatment is not constant, then it is larger for some people than for others. Perhaps the NSW training program changed the lives of some people and

had little or no effect on many others. David Salsburg [70] argued that common statistical procedures were poorly equipped to detect uncommon but large treatment effects, but he argued that such effects were often important; see also [71]. William Conover and David Salsburg [9] proposed a type of mixture model in which a small fraction of the population responds strongly to treatment, so $r_{Tij} - r_{Cij}$ is large, while the rest of the population is unaffected by treatment, so $r_{Tij} - r_{Cij} = 0$. Under this model, they derived the locally most powerful rank test. The ranks they obtained are difficult to interpret, so they do not lend themselves to use with attributable effects. It turns out, however, that for large I, after rescaling, their ranks are virtually the same as the ranks (2.7) that Stephenson [77] proposed on the basis of other considerations. As has been seen, Stephenson's ranks have a straightforward interpretation in terms of attributable effects. This subject is developed in detail in Chap. 17 and [64]. In short, the methods described in this section are well adapted to the study of uncommon but dramatic responses to treatment of the type that seem to have occurred in the NSW experiment.

2.6 Internal and External Validity

In Fisher's theory of randomized experiments, the inference is from the responses the $2I$ subjects did exhibit under the treatments they did receive, and it is to the $2I$ unobserved responses these same individuals would have exhibited had they received the alternative treatment. In brief, the inference concerns the unobservable causal effects $r_{Tij} - r_{Cij}$ for the $2I$ individuals in the experiment. See, for instance, Welch [83].

A distinction is often made between "internal" and "external" validity [72]. A randomized experiment is said to have a high level of "internal validity" in the sense that randomization provides a strong or "reasoned" basis for inference about the effects of the treatment, $r_{Tij} - r_{Cij}$, on the $2I$ individuals in the experiment. Causal inference is challenging even if one were only interested in these $2I$ individuals, and randomization meets that specific challenge. Often, one is interested not just in these $2I$ people, but in people generally, people on other continents, people who will be born in the future. "External" validity refers to the effects of the treatment on people not included in the experiment.

For instance, suppose that four large hospitals, say in Europe and the USA, cooperated in a randomized trial of two forms of surgery, with 60 day mortality as the outcome. Properly conducted, the trial would provide a sound basis for inference about the effects of the two forms of surgery on the $2I$ patients in this one clinical trial. One might plausibly think that similar effects would be found at other large hospitals in Europe and the USA, Canada, Japan, and other countries, but nothing in a statistical analysis based on four hospitals provides a warrant for this plausible

extrapolation. The extrapolation sounds plausible because of what we know about surgeons and hospitals in these countries, not because data from four cooperating hospitals provides a sound basis for extrapolation. Would similar effects be found at the small ambulatory surgical centers that are increasingly common in the USA and elsewhere? Will the same effects be found at large hospitals 10 years from now as other technologies improve? Statistical analysis of a clinical trial at four large hospitals provides no warrant for these extrapolations, although surgeons may be able to offer an informed guess. These questions refer to the degree of external validity of the clinical trial.

The common view [72], which I share, is that internal validity comes first. If you do not know the effects of the treatment on the individuals in your study, you are not well-positioned to infer the effects on individuals you did not study who live in circumstances you did not study. Randomization addresses internal validity. In actual practice, external validity is often addressed by comparing the results of several internally valid studies conducted in different circumstances at different times. These issues also arise in observational studies, and so will be discussed at various points later on.

2.7 Summary

The goal of causal inference in the NSW experiment is not to say that the earnings of treated and control men differed systematically, but rather that the treatment was the cause of the difference, so that if the treatment had been withheld from the treated men, there would have been no systematic difference between the two groups. Many statistical procedures can recognize a systematic difference, providing the sample size is sufficiently large. Randomization plays a key role in experiments that study the effects actually caused by a treatment.

The current chapter has done several things. It has introduced many elements that randomized experiments and observational studies have in common, such as treatment effects, treatment assignments, and pretreatment covariates unaffected by treatment. The chapter has discussed the large role played by random treatment assignment in experiments in justifying, or warranting, or being the "reasoned basis" for inference about effects caused by treatment. Without random assignment, in observational studies, this justification or warrant or reasoned basis is absent. Therefore, the chapter has framed the problem faced in an observational study, where the goal is again to draw inferences about the effects caused by treatment, but treatments were not assigned at random [8].

2.8 Appendix: Randomization Distribution of m-Statistics

2.8.1 Giving Observations Controlled Weight Using a ψ Function

Sections 2.3.2 and 2.3.3 were similar in structure, even though the sample mean $\overline{Y}$ and Wilcoxon's signed rank statistic are not similar as statistics. It was suggested in Sect. 2.3.3 that randomization inference is very general and virtually any statistic could be handled in a parallel manner. This is illustrated using another statistic, an m-test, that is, a test associated with Peter Huber's [37] m-estimates. The discussion here derives from a paper by J. S. Maritz [47]; see also [48, 63, 66]. It should be emphasized that the statistics used in m-tests differ in small but consequential ways from the statistics used in m-estimation, and the description that follows is for m-tests.[12]

As in Huber's m-estimation of a location parameter, Maritz's test statistic involves a function $\psi(\cdot)$ that gives weights to scaled observations. Rather than "reject outliers" from a statistical analysis, a suitable ψ-function gradually reduces the weight of observations as they become increasingly extreme. Because we want an observation that is high by 3 units to pull up as strongly as an observation that is low by 3 units pulls down, the function $\psi(\cdot)$ is required to be an odd function, meaning $\psi(-y) = -\psi(y)$ for every $y \geq 0$, so in particular $\psi(-3) = -\psi(3)$ and $\psi(0) = 0$. For testing and confidence intervals, it is wise to insist that $\psi(\cdot)$ be monotone increasing, so $y < y'$ implies $\psi(y) \leq \psi(y')$. One such function, similar to a trimmed mean, was proposed by Huber; it has $\psi(y) = \max\{-1, \min(y, 1)\}$, so that $\psi(y) = -1$ for $y < -1$, $\psi(y) = y$ for $-1 \leq y \leq 1$, and $\psi(y) = 1$ for $y > 1$. In effect, the mean $\overline{Y}$ has a $\psi(\cdot)$ function of $\psi(y) = y$.

Consider using the treated-minus-control differences in paired responses, $Y_i = (Z_{i1} - Z_{i2})(R_{i1} - R_{i2})$, in testing the hypothesis of an additive treatment effect, $H_0 : r_{Tij} = \tau_0 + r_{Cij}$, for all ij, where τ_0 is a specified number. When $\tau_0 = 0$, this is Fisher's sharp null hypothesis of no treatment effect, and this is the case presented in detail in Table 2.12. As seen in Sect. 2.4, if H_0 were true, then the treated-minus-control difference in responses in pair i is $Y_i = \tau_0 + (Z_{i1} - Z_{i2})(r_{Ci1} - r_{Ci2})$. In particular, if H_0 were true, $Y_i - \tau_0$ would be $\pm(r_{Ci1} - r_{Ci2})$ and $|Y_i - \tau_0| =$

[12]The difference involves the scaling factor. In estimation, the scaling factor might be the median absolute deviation from the sample median. In testing, the scaling factor might be the median absolute deviation from the hypothesized median, which for testing no effect is zero. In testing a hypothesis, one can presume the hypothesis is true for the purpose of testing it. By using the hypothesized median rather than the sample median, the scaling factor becomes fixed as the treatment assignments vary—as illustrated later—with the consequence that Maritz's statistic has a null randomization distribution that is the sum of independent bounded random variables, and this greatly simplifies the creation of a large sample approximation to the null randomization distribution. This small difference has various consequences that should be considered before designing a new m-test; see [47, 63] for discussion.

Table 2.12 The possible treated-minus-control differences in adjusted responses $Y_i - \tau_0 = (Z_{i1} - Z_{i2})(r_{Ci1} - r_{Ci2})$ and Huber–Maritz m-statistic, T^*, under the null hypothesis of an additive treatment effect, $H_0 : r_{Tij} = r_{Cij} + \tau_0$, for $\tau_0 = 0$ for the small version of the NSW experiment with $I = 5$ pairs

Label	Pair 1	Pair 2	Pair 3	Pair 4	Pair 5	Huber-Maritz
	$Y_1 - \tau_0$	$Y_2 - \tau_0$	$Y_3 - \tau_0$	$Y_4 - \tau_0$	$Y_5 - \tau_0$	T^*
1	1456	3988	−45	−2147	3173	1.66
2	1456	3988	−45	−2147	−3173	−0.34
3	1456	3988	−45	2147	3173	3.66
4	1456	3988	−45	2147	−3173	1.66
5	1456	3988	45	−2147	3173	1.70
6	1456	3988	45	−2147	−3173	−0.30
7	1456	3988	45	2147	3173	3.70
8	1456	3988	45	2147	−3173	1.70
9	1456	−3988	−45	−2147	3173	−0.34
10	1456	−3988	−45	−2147	−3173	−2.34
11	1456	−3988	−45	2147	3173	1.66
12	1456	−3988	−45	2147	−3173	−0.34
13	1456	−3988	45	−2147	3173	−0.30
14	1456	−3988	45	−2147	−3173	−2.30
15	1456	−3988	45	2147	3173	1.70
16	1456	−3988	45	2147	−3173	−0.30
17	−1456	3988	−45	−2147	3173	0.30
18	−1456	3988	−45	−2147	−3173	−1.70
19	−1456	3988	−45	2147	3173	2.30
20	−1456	3988	−45	2147	−3173	0.30
21	−1456	3988	45	−2147	3173	0.34
22	−1456	3988	45	−2147	−3173	−1.66
23	−1456	3988	45	2147	3173	2.34
24	−1456	3988	45	2147	−3173	0.34
25	−1456	−3988	−45	−2147	3173	−1.70
26	−1456	−3988	−45	−2147	−3173	−3.70
27	−1456	−3988	−45	2147	3173	0.30
28	−1456	−3988	−45	2147	−3173	−1.70
29	−1456	−3988	45	−2147	3173	−1.66
30	−1456	−3988	45	−2147	−3173	−3.66
31	−1456	−3988	45	2147	3173	0.34
32	−1456	−3988	45	2147	−3173	−1.66
All	Absolute Differences $\lvert Y_i - \tau_0 \rvert$					
1–32	1456	3988	45	2147	3173	
All	Scaled Absolute Differences $\lvert Y_i - \tau_0 \rvert / s_{\tau_0}$					
1–32	0.68	1.86	0.02	1.00	1.48	
All	Scored, Scaled Absolute Differences $\psi(\lvert Y_i - \tau_0 \rvert / s_{\tau_0})$					
1–32	0.68	1.00	0.02	1.00	1.00	

$|r_{Ci1} - r_{Ci2}|$ would be fixed[13] not varying with the randomly assigned treatment Z_{ij}. This is visible for $\tau_0 = 0$ at the bottom of Table 2.12: assuming $H_0 : r_{Tij} = \tau_0 + r_{Cij}$ with $\tau_0 = 0$ for the purpose of testing it, the $|Y_i - \tau_0|$ are the same for all $|\mathscr{Z}| = 32$ treatment assignments in $\mathscr{Z}$.

2.8.2 Scaling

The observations are scaled before the $\psi(\cdot)$ function is calculated. The scaling factor s_{τ_0} used by Maritz's [47] statistic is a specific quantile of the I absolute differences, $|Y_i - \tau_0|$, most commonly the median, $s_{\tau_0} = \text{median}\,|Y_i - \tau_0|$. When used with Huber's $\psi(y) = \max\{-1, \min(y, 1)\}$, scaling by $s_{\tau_0} = \text{median}\,|Y_i - \tau_0|$ trims half the observations and is analogous to a midmean. If a larger quantile of the $|Y_i - \tau_0|$ is used for scaling, say the 75th or the 90th quantile rather than the median, then there is less trimming. If the maximum $|Y_i - \tau_0|$ is used for scaling with Huber's $\psi(y) = \max\{-1, \min(y, 1)\}$, then Maritz's statistic becomes the sample mean and it reproduces the randomization inference in Sect. 2.3.2.

Because s_{τ_0} is calculated from $|Y_i - \tau_0|$, if $H_0 : r_{Tij} = \tau_0 + r_{Cij}$, were true, then s_{τ_0} would have the same value for each of the 32 treatment assignments, or $s_{\tau_0} = 2147$ in Table 2.12. In the example in Table 2.12, the calculations continue with Huber's function $\psi(y) = \max\{-1, \min(y, 1)\}$, specifically the scaled absolute responses, $|Y_i - \tau_0| / s_{\tau_0}$, and the scored, scaled absolute responses, $\psi\left(|Y_i - \tau_0| / s_{\tau_0}\right)$. Notice that $\psi\left(|Y_i - \tau_0| / s_{\tau_0}\right)$ takes the same value for all 32 treatment assignments.

2.8.3 Randomization Test of a Hypothesized Additive Effect, $H_0 : r_{Tij} = r_{Cij} + \tau_0$

For testing the hypothesis of an additive treatment effect, $H_0 : r_{Tij} = \tau_0 + r_{Cij}$, Maritz's [47] m-test statistic is $T^* = \sum_{i=1}^{I} \psi\left\{(Y_i - \tau_0) / s_{\tau_0}\right\}$. Recall that $\psi(\cdot)$ is an odd function, $\psi(-y) = -\psi(y)$ for $y \geq 0$, so $\psi(y) = \text{sign}(y) \cdot \psi(|y|)$ where $\text{sign}(y) = 1$ if $y > 0$, $\text{sign}(y) = 0$ if $y = 0$, $\text{sign}(y) = -1$ if $y < 0$. If $H_0 : r_{Tij} = \tau_0 + r_{Cij}$ were true, then $Y_i - \tau_0 = (Z_{i1} - Z_{i2})(r_{Ci1} - r_{Ci2})$, so $T^* = \sum_{i=1}^{I} \text{sign}(Y_i - \tau_0) \cdot \psi\left(|Y_i - \tau_0| / s_{\tau_0}\right)$ or equivalently

$$T^* = \sum_{i=1}^{I} \text{sign}\{(Z_{i1} - Z_{i2})(r_{Ci1} - r_{Ci2})\} \cdot \psi\left(|r_{Ci1} - r_{Ci2}| / s_{\tau_0}\right). \tag{2.14}$$

[13] That is, $|Y_i - \tau_0| = |r_{Ci1} - r_{Ci2}|$ is determined by $\mathscr{F}$.

In words, if $H_0 : r_{Tij} = \tau_0 + r_{Cij}$ were true, T^* would be the sum of I independent quantities that take the fixed values $\psi\left(|r_{Ci1} - r_{Ci2}|/s_{\tau_0}\right)$ and $-\psi\left(|r_{Ci1} - r_{Ci2}|/s_{\tau_0}\right)$ each with probability $\frac{1}{2}$.

For instance, the observed value of T^* is 1.66 in the first row of Table 2.12; that is, $T^* = 1.66 = 0.68+1.00+(-0.02)+(-1.00)+1.00$. Ten of the 32 equally probable treatment assignments $\mathbf{z} \in \mathscr{Z}$ yield values of T^* that are at least 1.66, so the one-sided P-value for testing $H_0 : r_{Tij} = \tau_0 + r_{Cij}$ with $\tau_0 = 0$ is $\frac{10}{32} = 0.3125 = \Pr(T^* \geq 1.66|\ \mathscr{F}, \mathscr{Z})$. The two-sided P-value is $\frac{20}{32} = 0.625$ which equals both $2 \cdot \Pr(T^* \geq 1.66|\ \mathscr{F}, \mathscr{Z})$ and $\Pr(T^* \leq -1.66|\ \mathscr{F}, \mathscr{Z}) + \Pr(T^* \geq 1.66|\ \mathscr{F}, \mathscr{Z})$ because the null distribution of T^* given $\mathscr{F}, \mathscr{Z}$ is symmetric about zero.

Table 2.12 calculates the exact null distribution of T^*, but a large sample approximation using the central limit theorem is easy to obtain as $I \to \infty$. The null distribution of T^* is the distribution of the sum of I independent terms taking the fixed values $\pm\psi\left(|r_{Ci1} - r_{Ci2}|/s_{\tau_0}\right)$ each with probability $\frac{1}{2}$. Therefore, in a randomized experiment under the null hypothesis, $E(T^*|\ \mathscr{F}, \mathscr{Z}) = 0$ and $\mathrm{var}(T^*|\ \mathscr{F}, \mathscr{Z}) = \sum_{i=1}^{I}\left\{\psi\left(|r_{Ci1} - r_{Ci2}|/s_{\tau_0}\right)\right\}^2$. With a bounded $\psi(\cdot)$ function, such as Huber's $\psi(y) = \max\{-1, \min(y, 1)\}$, as $I \to \infty$, the null distribution of $T^*/\sqrt{\mathrm{var}(T^*|\ \mathscr{F}, \mathscr{Z})}$ converges in distribution to the standard Normal distribution provided $\Pr\left(s_{\tau_0} > 0\right) \to 1$; see [47].

The discussion in this section has focused on matched pairs, but essentially the same reasoning may be applied when matching with multiple controls [63].

2.8.4 m-Tests in the NSW Experiment

As seen in Sect. 2.5, in the NSW experiment, much of the treatment effect is in the tail, so less trimming is better. In testing the hypothesis of no effect, $H_0 : \tau_0 = 0$, scaling by $s_{\tau_0} = \text{median}\,|Y_i - \tau_0|$ trims half the observations and yields a two-sided P-value of 0.040, while scaling by the 90th percentile of $|Y_i - \tau_0|$ yields a two-sided P-value of 0.0052. Using the 90th percentile, the 95% confidence interval for an additive effect is \$475 to \$2788.

2.9 Further Reading

Randomized experimentation is due to Sir Ronald Fisher [22], whose book of 1935 is of continuing interest. David Cox and Nancy Reid [15] present a modern discussion of the theory of randomized experiments. For randomized clinical trials in medicine, see [27, 52], and for experiments in the social sciences, see [7, 10, 28]. Among books that discuss permutation and randomization inferences, Erich Lehmann's [43] book *Nonparametrics* is especially careful in distinguishing properties that follow from random assignment and properties that involve sampling

a population. For causal effects as comparisons of potential outcomes under alternative treatments, see the papers by Jerzy Neyman [49], B.L. Welch [83], and Donald Rubin [68]. For a critical view, see A.P. Dawid [16]. Donald Rubin's paper [68] is the first parallel, formal discussion of causal effects in randomized and observational studies. For three important illustrations of the enhanced clarity provided by this approach, see [25, 26, 69]. Fisher [22], Welch [83], Kempthorne [40], Wilk [85], Cox [13], Robinson [55], and many others viewed sampling models as approximations to randomization inferences. Randomization inference is, of course, not limited to matched pairs [43, 59] and it may be used in conjunction with general forms of covariance adjustment [23, 60, 74, 80, 81].

What is the relationship between permutation inference in a randomized experiment and permutation inference in a sample from an infinite population? Several sections of Chapter 5 of Lehmann and Romano's [44, §5.8-§5.12] textbook develop this topic in a satisfying and beautiful way. Using the complete sufficiency of the order statistics, they show that a test can have the correct level for all continuous distributions *only* if it is a permutation test [44, Theorem 5.8.1]. In their formulation, the confidence interval formed by inverting a permutation test is the same interval whether it is for an additive effect on individuals in a randomized experiment or a shift in probability distributions in a sample from an infinite population.

In labor economics, the National Supported Work experiment became a small laboratory for methods for observational studies when Robert LaLonde [41] set aside the randomized control group and replaced it with a nonrandomized control group drawn from a survey; see [17, 30, 31, 41, 75] for discussion.

2.10 Software

Randomization inference using Wilcoxon's signed rank statistic is widely implemented. See, for instance, the function `wilcox.test` in the `stats` package in `R`, or the `senWilcox` and `senWilcoxExact` functions in the `DOS2` package.

Stephenson's test [77] in Sect. 2.5 is implemented in the `senU` function in the `DOS2` package [64, 65].

Randomization inference using the mean is a special case of randomization inference using an m-statistic. Randomization inference with an m-statistic [47, 63] in Sect. 2.8 is implemented in the `senm` and `senmCI` functions of the `R` package `sensitivitymult`. See also packages `sensitivitymv`, `sensitivitymw`, and `sensitivityfull`. These packages are discussed in [67].

2.11 Data

The $I = 185$ pairs from the National Supported Work experiment are in the `DOS2` package in `R` as `NSW`.

Bibliography

1. Andrews, D.F., Bickel, P.J., Hampel, F.R., Huber, P.J., Rogers, W.H., Tukey, J.W.: Robust Estimates of Location. Princeton University Press, Princeton, NJ (1972)
2. Athey, S., Eckles, D., Imbens, G.W.:Exact p-values for network interference. J. Am. Stat. Assoc. **113**, 230–240 (2018)
3. Baiocchi, M., Small, D.S., Lorch, S., Rosenbaum, P.R.: indexBaiocchi, M. Building a stronger instrument in an observational study of perinatal care for premature infants. J. Am. Stat. Assoc. **105**, 1285–1296 (2010)
4. Basse, G., Feller, A.: Analyzing two-stage experiments in the presence of interference. J. Am. Stat. Assoc. **113**, 41–55 (2018)
5. Basse, G.W., Feller, A., Toulis, P.: Randomization tests of causal effects under interference. Biometrika **106**, 487–494 (2019)
6. Bennett, D.M.: Confidence limits for a ratio using Wilcoxon's signed rank test. Biometrics **21**, 231–234 (1965)
7. Boruch, R.: Randomized Experiments for Planning and Evaluation. Sage, Thousand Oaks, CA (1997)
8. Cochran, W.G.: The planning of observational studies of human populations (with Discussion). J. R. Stat. Soc. A **128**, 234–265 (1965)
9. Conover, W.J., Salsburg, D.S.: Locally most powerful tests for detecting treatment effects when only a subset of patients can be expected to 'respond' to treatment. Biometrics **44**, 189–196 (1988)
10. Cook, T.D., Shadish, W.R.: Social experiments: some developments over the past fifteen years. Annu. Rev. Psychol. **45**, 545–580 (1994)
11. Couch, K.A.: New evidence on the long-term effects of employment training programs. J. Labor Econ. **10**, 380–388 (1992)
12. Cox, D.R.: Planning of Experiments. Wiley, New York (1958)
13. Cox, D.R.: The interpretation of the effects of non-additivity in the Latin square. Biometrika **45**, 69–73 (1958)
14. Cox, D.R.: The role of significance tests. Scand. J. Stat. **4**, 49–62 (1977)
15. Cox, D.R., Reid, N.: The Theory of the Design of Experiments. Chapman and Hall/CRC, New York (2000)
16. Dawid, A.P.: Causal inference without counterfactuals. J. Am. Stat. Assoc. **95**, 407–424 (2000)
17. Dehejia, R.H., Wahba, W.: Causal effects in nonexperimental studies: reevaluating the evaluation of training programs. J. Am. Stat. Assoc. **94**, 1053–1062 (1999)
18. Ding, P., Feller, A., Miratrix, L.: Randomization inference for treatment effect variation. J R. Stat. Soc. B **78**, 655–671 (2016)
19. Ding, P., Li, X., Miratrix, L.W.: Bridging finite and super population causal inference. J. Causal Infer. 20160027 (2017). https://doi.org/10.1515/jci-2016-0027
20. Eckles, D., Karrer, B., Ugander, J.: Design and analysis of experiments in networks: reducing bias from interference. J. Causal Infer. **5**(1) (2017). https://doi.org/10.1515/jci-2015-0021
21. Efron, B.: Student's t-test under symmetry conditions. J. Am. Stat. Assoc. **64**, 1278–1302 (1969)
22. Fisher, R.A.: Design of Experiments. Oliver and Boyd, Edinburgh (1935)
23. Fogarty, C.B.: Regression-assisted inference for the average treatment effect in paired experiments. Biometrika **105**, 994–1000 (2018)
24. Fogarty, C.B., Shi, P., Mikkelsen, M.E., Small, D.S.: Randomization inference and sensitivity analysis for composite null hypotheses with binary outcomes in matched observational studies. J. Am. Stat. Assoc. **112**, 321–331 (2017)
25. Frangakis, C.E., Rubin, D.B.: Principal stratification in causal inference. Biometrics **58**, 21–29 (2002)
26. Freedman, D.A.: Randomization does not justify logistic regression. Stat. Sci. **23**, 237–249 (2008)

27. Friedman, L.M., DeMets, D.L., Furberg, C.D.: Fundamentals of Clinical Trials. Springer, New York (1998)
28. Gerber, A.S., Green, D.P.: Field Experiments: Design, Analysis, and Interpretation. Norton, New York (2012)
29. Greevy, R., Lu, B., Silber, J.H., Rosenbaum, P.R.: Optimal matching before randomization. Biostatistics **5**, 263–275 (2004)
30. Hämäläinen, K., Uusitalo, R., Vuori, J.: Varying biases in matching estimates: evidence from two randomized job search training experiments. Labour Econ. **15**, 604–618 (2008)
31. Heckman, J.J., Hotz, V.J.: Choosing among alternative nonexperimental methods for estimating the impact of social programs: the case of manpower training (with Discussion). J. Am. Stat. Assoc. **84**, 862–874 (1989)
32. Hodges, J.L., Lehmann, E.L.: Rank methods for combination of independent experiments in analysis of variance. Ann. Math. Stat. **33**, 482–497 (1962)
33. Hodges, J.L., Lehmann, E.L.: Estimates of location based on ranks. Ann. Math. Stat. **34**, 598–611 (1963)
34. Hoeffding, W.: A class of statistics with asymptotically normal distributions. Ann. Math. Stat. **19**, 293–325 (1948)
35. Holland, P.W.: Statistics and causal inference. J. Am. Stat. Assoc. **81**, 945–960 (1986)
36. Hollister, R., Kemper, P., Maynard, R. (eds.): The National Supported Work Demonstration. University of Wisconsin Press, Madison (1984)
37. Huber, P.J.: Robust estimation of a location parameter. Ann. Math. Stat. **35**, 73–101 (1964)
38. Hudgens, M.G., Halloran, M.E.:Toward causal inference with interference. J. Am. Stat. Assoc. **482**, 832–842 (2008)
39. Jurečková, J., Sen, P.K.: Robust Statistical Procedures. Wiley, New York (1996)
40. Kempthorne, O.: Design and Analysis of Experiments. Wiley, New York (1952)
41. LaLonde, R.J.: Evaluating the econometric evaluations of training programs with experimental data. Am. Econ. Rev. **76**, 604–620 (1986)
42. LaLonde, R.J.: The promise of public sector-sponsored training programs. J. Econ. Perspect. **9**, 149–168 (1995)
43. Lehmann, E.L.: Nonparametrics. Holden-Day, San Francisco (1975)
44. Lehmann, E.L., Romano, J.P.: Testing Statistical Hypotheses, 3rd edn. Springer, New York (2006)
45. Li, X., Ding, P.: Exact confidence intervals for the average causal effect on a binary outcome. Stat. Med. **35**, 957–960 (2016)
46. Li, X., Ding, P.: General forms of finite population central limit theorems with applications to causal inference. J. Am. Stat. Assoc. **112**, 1759–1769 (2017)
47. Maritz, J.S.: A note on exact robust confidence intervals for location. Biometrika **66**, 163–166 (1979)
48. Maritz, J.S.: Distribution-Free Statistical Methods. Chapman and Hall, London (1995)
49. Neyman, J.: On the application of probability theory to agricultural experiments: essay on principles, Section 9. In Polish, but reprinted in English with Discussion by T. Speed and D.B. Rubin in Statist Sci **5**, 463–480 (1923, reprinted 1990)
50. Nichols, T.E., Holmes, A.P.: Nonparametric permutation tests for functional neuroimaging. Human Brain Mapping **15**, 1–25
51. Pagano, M., Tritchler, D.: On obtaining permutation distributions in polynomial time. J. Am. Stat. Assoc. **78**, 435–440 (1983)
52. Piantadosi, S.: Clinical Trials. Wiley, New York (2005)
53. Pratt, J.W., Gibbons, J.D.: Concepts of Nonparametric Theory. Springer, New York (1981)
54. Rigdon, J., Hudgens, M.G.: Randomization inference for treatment effects on a binary outcome. Stat. Med. **34**, 924–935 (2015)
55. Robinson, J.: The large sample power of permutation tests for randomization models. Ann. Stat. **1**, 291–296 (1973)

56. Rosenbaum, P.R.: The consequences of adjustment for a concomitant variable that has been affected by the treatment. J. R. Stat. Soc. A **147**, 656–666 (1984)
57. Rosenbaum, P.R.: Effects attributable to treatment: inference in experiments and observational studies with a discrete pivot. Biometrika **88**, 219–231 (2001)
58. Rosenbaum, P.R.: Attributing effects to treatment in matched observational studies. J. Am. Stat. Assoc. **97**, 183–192 (2002)
59. Rosenbaum, P.R.: Observational Studies, 2nd edn. Springer, New York (2002)
60. Rosenbaum, P.R.: Covariance adjustment in randomized experiments and observational studies (with Discussion). Stat. Sci. **17**, 286–327 (2002)
61. Rosenbaum, P.R.: Exact confidence intervals for nonconstant effects by inverting the signed rank test. Am. Stat. **57**, 132–138 (2003)
62. Rosenbaum, P.R.: Interference between units in randomized experiments. J. Am. Stat. Assoc. **102**, 191–200 (2007)
63. Rosenbaum, P.R.: Sensitivity analysis for m-estimates, tests, and confidence intervals in matched observational studies. Biometrics **63**, 456–464 (2007)
64. Rosenbaum, P.R.: Confidence intervals for uncommon but dramatic responses to treatment. Biometrics **63**, 1164–1171 (2007)
65. Rosenbaum, P.R.: A new U-statistic with superior design sensitivity in matched observational studies. Biometrics **67**, 1017–1027 (2011)
66. Rosenbaum, P.R.: Impact of multiple matched controls on design sensitivity in observational studies. Biometrics **69**, 118–127 (2013)
67. Rosenbaum, P.R.: Two R packages for sensitivity analysis in observational studies. Observ. Stud. **1**, 1–17 (2015)
68. Rubin, D.B.: Estimating causal effects of treatments in randomized and nonrandomized studies. J. Educ. Psychol. **66**, 688–701 (1974)
69. Rubin, D.B., Stuart, E.A., Zanutto, E.L.: A potential outcomes view of value-added assessment in education. J. Educ. Behav. Stat. **29**, 103–116 (2004)
70. Salsburg, D.S.: Alternative hypotheses for the effects of drugs in small-scale clinical studies. Biometrics **42**, 671–674 (1986)
71. Salsburg, D.S.: The Use of Restricted Significance Tests in Clinical Trials. Springer, New York (1992)
72. Shadish, W.R., Cook, T.D., Campbell, D.T.: Experimental and Quasi-Experimental Designs for Generalized Causal Inference. Houghton-Mifflin, Boston (2002)
73. Shaffer, J.P.: Bidirectional unbiased procedures. J. Am. Stat. Assoc. **69**, 437–439 (1974)
74. Small, D., Ten Have, T.R., Rosenbaum, P.R.: Randomization inference in a group-randomized trial of treatments for depression: covariate adjustment, noncompliance and quantile effects. J. Am. Stat. Assoc. **103**, 271–279 (2007)
75. Smith, J.A., Todd, P.E.: Does matching overcome LaLonde's critique of nonexperimental estimators? J Econometrics **125**, 305–353 (2005)
76. Sobel, M.E.: What do randomized studies of housing mobility demonstrate? Causal inference in the face of interference. J. Am. Stat. Assoc. **476**, 1398–1407 (2006)
77. Stephenson, W.R.: A general class of one-sample nonparametric test statistics based on subsamples. J. Am. Stat. Assoc. **76**, 960–966 (1981)
78. Tchetgen Tchetgen, E.J., VanderWeele, T.J.: On causal inference in the presence of interference. Stat. Meth. Med. Res. **21**, 55–75 (2012)
79. Tobin, J.: Estimation of relationships for limited dependent variables. Econometrica **26**, 24–36 (1958)
80. Wager, S., Athey, S.: Estimation and inference of heterogeneous treatment effects using random forests. J. Am. Stat. Assoc. **113**, 1228–1242 (2018)
81. Wager, S., Du, W., Taylor, J., Tibshirani, R.J.:High-dimensional regression adjustments in randomized experiments. Proc. Nat. Acad. Sci. **113**, 12673–12678 (2016)

82. Weiss, L.: A note on confidence sets for random variables. Ann. Math. Stat. **26**, 142–144 (1955)
83. Welch, B.L.: On the z-test in randomized blocks and Latin squares. Biometrika **29**, 21–52 (1937)
84. Wilcoxon, F.: Individual comparisons by ranking methods. Biometrics **1**, 80–83 (1945)
85. Wilk, M.B.: The randomization analysis of a generalized randomized block design. Biometrika **42**, 70–79 (1955)

Chapter 3
Two Simple Models for Observational Studies

Abstract Observational studies differ from experiments in that randomization is not used to assign treatments. How were treatments assigned? This chapter introduces two simple models for treatment assignment in observational studies. The first model is useful but naïve: it says that people who look comparable are comparable. The second model speaks to a central concern in observational studies: people who look comparable in the observed data may not actually be comparable; they may differ in ways we did not observe.

3.1 The Population Before Matching

In the population before matching, there are L subjects, $\ell = 1, 2, \ldots, L$. Like the subjects in the randomized experiment in Chap. 2, each of these L subjects has an observed covariate $\mathbf{x}_\ell$, an unobserved covariate u_ℓ, an indicator of treatment assignment, Z_ℓ, where $Z_\ell = 1$ if the subject receives treatment or $Z_\ell = 0$ if the subject receives the control, a potential response to treatment, $r_{T\ell}$, which is seen if the subject receives treatment, $Z_\ell = 1$, a potential response to control, $r_{C\ell}$, which is seen if the subject receives the control, $Z_\ell = 0$, and an observed response, $R_\ell = Z_\ell\, r_{T\ell} + (1 - Z_\ell)\, r_{C\ell}$. Now, however, treatments Z_ℓ are not assigned by the equitable flip of a fair coin.

In the population before matching, we imagine that subject ℓ received treatment with probability π_ℓ, independently of other subjects, where π_ℓ may vary from one person to the next and is not known. More precisely,

$$\pi_\ell = \Pr(Z_\ell = 1 \mid r_{T\ell},\, r_{C\ell},\, \mathbf{x}_\ell,\, u_\ell)\,, \tag{3.1}$$

$$\begin{aligned}&\Pr(Z_1 = z_1, \ldots, Z_L = z_L \mid r_{T1},\, r_{C1},\, \mathbf{x}_1,\, u_1, \ldots, r_{TL},\, r_{CL},\, \mathbf{x}_L,\, u_L)\\ &= \pi_1^{z_1}\,(1-\pi_1)^{1-z_1} \cdots \pi_L^{z_L}\,(1-\pi_L)^{1-z_L} = \Pi_{\ell=1}^{L}\, \pi_\ell^{z_\ell}\,(1-\pi_\ell)^{1-z_\ell}\,.\end{aligned}$$

It is natural to ask whether very little or a great deal has just been assumed. One might reasonably worry if a great deal has just been assumed, for perhaps the

© Springer Nature Switzerland AG 2020

P. R. Rosenbaum, *Design of Observational Studies*, Springer Series in Statistics,
https://doi.org/10.1007/978-3-030-46405-9_3

central problems in observational studies have just been assumed away. Addressing difficult problems by assuming they are not there is one of the less reliable tactics, so it is reasonable to seek some assurance that this has not just happened. Indeed, it has not. Actually, because of the delicate nature of unobserved variables, nothing at all has been assumed: there is always an unobserved u_ℓ that makes all of this true, and true in a fairly trivial way. Specifically, if we set $u_\ell = Z_\ell$, then $\pi_\ell = \Pr(Z_\ell = 1 \mid r_{T\ell}, r_{C\ell}, \mathbf{x}_\ell, u_\ell)$ is $\Pr(Z_\ell = 1 \mid r_{T\ell}, r_{C\ell}, \mathbf{x}_\ell, Z_\ell)$ which is, of course, $\pi_\ell = 1$ if $Z_\ell = 1$ and $\pi_\ell = 0$ if $Z_\ell = 0$; moreover, with π_ℓ so defined, $\Pi_{\ell=1}^{L} \pi_\ell^{z_\ell} (1 - \pi_\ell)^{1-z_\ell}$ is 1 for the observed treatment assignment and 0 for all other treatment assignments, and

$$\begin{aligned}&\Pr(Z_1 = z_1, \ldots, Z_L = z_L | r_{T1}, r_{C1}, \mathbf{x}_1, u_1, \ldots, r_{TL}, r_{CL}, \mathbf{x}_L, u_L)\\ &= \Pr(Z_1 = z_1, \ldots, Z_L = z_L | r_{T1}, r_{C1}, \mathbf{x}_1, Z_1, \ldots, r_{TL}, r_{CL}, \mathbf{x}_L, Z_L)\end{aligned}$$

is also 1 for the observed treatment assignment and 0 for all other treatment assignments; so it is all true, in a trivial way, with $u_\ell = Z_\ell$. Until something is done to restrict the behavior of the unobserved covariate u_ℓ, nothing has been assumed: (3.1) is a representation, not a model, because there is always some u_ℓ that makes (3.1) true. A representation, such as (3.1), is a manner of speaking; it does not become a model until it says something that may be false.

3.2 The Ideal Matching

Imagine that we could find two subjects, say k and ℓ, such that exactly one was treated, $Z_k + Z_\ell = 1$, but they had the same probability of treatment, $\pi_k = \pi_\ell$, and we made those two subjects into a matched pair.[1] Obviously, it is something of a challenge to create this matched pair because we do not observe u_k and u_ℓ, and we observe either r_{Tk} or r_{Ck} but not both, and either $r_{T\ell}$ or $r_{C\ell}$ but not both. It is a challenge because we cannot create this pair by matching on observable quantities. We could create this pair by matching on the observable $\mathbf{x}$, so $\mathbf{x}_k = \mathbf{x}_\ell$, and then flipping a fair coin to determine (Z_k, Z_ℓ), assigning one member of the pair to treatment, the other to control; that is, we could create this pair by conducting a randomized paired experiment, as in Chap. 2. Indeed, with the aid of random assignment, matching on $\mathbf{x}$ may be prudent [39], but is not necessary to achieve $\pi_k = \pi_\ell$ because the π's are created by the experimenter. If randomization is infeasible or unethical, then how might one attempt to find a treated–control pair such that the paired units have the same probability of treatment, $\pi_k = \pi_\ell$?

[1]The alert reader will notice that the supposition that there could be two subjects, k and ℓ, with $Z_k + Z_\ell = 1$ but $\pi_k = \pi_\ell$ is already a step beyond the mere representation in (3.1), because it precludes setting $u_i = Z_i$ for all i and instead requires $0 < \pi_i < 1$ for at least two subjects.

Return for a moment to the study by Joshua Angrist and Victor Lavy [5] of class size and educational test performance, and in particular to the $I = 86$ pairs of Israeli schools in Sect. 1.3, one with between 31 and 40 students in the fifth grade ($Z = 0$), the other with between 41 and 50 students in the fifth grade ($Z = 1$), the pairs being matched for a covariate $\mathbf{x}$, namely the percentage of disadvantaged students. Strict adherence to Maimonides' rule would require the slightly larger fifth grade cohorts to be taught in two small classes and the slightly smaller fifth grade cohorts to be taught in one large class. As seen in Fig. 1.1, adherence to Maimonides' rule was imperfect but strict enough to produce a wide separation in typical class size. In their study, (r_T, r_C) are average test scores in fifth grade if the fifth grade cohort was a little larger ($Z = 1$) or a little smaller ($Z = 0$). What separates a school with a slightly smaller fifth grade cohort ($Z = 0$, 31–40 students) and a school with a slightly larger fifth grade cohort ($Z = 1$, 41–50 students)? Well, what separates them is the enrollment of a handful of students in grade 5. It seems reasonably plausible that whether or not a few more students enroll in the fifth grade is a relatively haphazard event, an event not strongly related to the average test performance (r_T, r_C) that the fifth grade would exhibit with a larger or smaller cohort. That is, building a study in the way that Angrist and Lavy did, it does seem reasonably plausible that π_k and π_ℓ are fairly close in the $I = 86$ matched pairs of two schools.

Properly understood, a "natural experiment" is an attempt to find in the world some rare circumstance such that a consequential treatment was handed to some people and denied to others for no particularly good reason at all, that is, haphazardly [4, 10, 42, 58, 69, 91, 112, 117, 124, 129]. The word "natural" has various connotations, but a "natural experiment" is a "wild experiment" not a "wholesome experiment," natural in the way that a tiger is natural, not in the way that oatmeal is natural. To express the same thought differently: to say that "it does seem reasonably plausible that π_k and π_ℓ are fairly close" is to say much less than "definitely, $\pi_k = \pi_\ell$ by the way treatments were randomly assigned." Haphazard is a far cry from randomized, a point given proper emphasis by Mark Rosenzweig and Ken Wolpin [112] and by Timothy Besley and Anne Case [10] in reviews of several natural experiments.

Nonetheless, it does seem reasonable to think that π_k and π_ℓ are fairly close for the 86 paired Israeli schools in Fig. 1.1. This might be implausible in some other context, say in a national survey in the USA where schools are funded by local governments, so that class size might be predicted by the wealth or poverty of the local community. Even in the Israeli schools, the claim that π_k and π_ℓ are plausibly close depended upon: (1) matching for the percentage of students who were disadvantaged, $\mathbf{x}_k$, and (2) defining Z_k in terms of cohort size rather than class size. In the pair of boxplots on the upper right of Fig. 1.1 depicting class size and cohort size, there is considerable adherence to Maimonides' rule, but there are a few deviations: in particular, two large fifth grade cohorts than strict adherence to the rule would have divided, and several small cohorts that were taught in smaller classes than strict adherence to the rule would have permitted. Small variations in the size of the fifth grade cohort are plausibly haphazard, but deviations from Maimonides'

rule may have been deliberate, considered responses to special circumstances, and hence not haphazard. As defined in Fig. 1.1, Z_k reflects the haphazard element, namely susceptibility to Maimonides' rule based on cohort size, rather than realized class size. Defining Z_k in this way is analogous to an "intention-to-treat" analysis (Sect. 1.2.6) in a randomized experiment.[2]

Suppose indeed we could find two subjects, k and ℓ, with the same probability of treatment, $\pi_k = \pi_\ell$, where the common value of π_k and π_ℓ is typically unknown. What would follow from this? From (3.1),

$$\Pr(Z_k = z_k,\ Z_\ell = z_\ell \,|\, r_{Tk},\ r_{Ck},\ \mathbf{x}_k,\ u_k,\ r_{T\ell},\ r_{C\ell},\ \mathbf{x}_\ell,\ u_\ell) \tag{3.2}$$

$$= \pi_k^{z_k}\ (1-\pi_k)^{1-z_k}\ \pi_\ell^{z_\ell}\ (1-\pi_\ell)^{1-z_\ell} \tag{3.3}$$

$$= \pi_\ell^{z_k+z_\ell}\ (1-\pi_\ell)^{(1-z_k)+(1-z_\ell)} \quad \text{because} \quad \pi_k = \pi_\ell, \tag{3.4}$$

which is, so far, of limited use, because π_ℓ is unknown. Suppose, however, that we found two subjects, k and ℓ, with the same probability of treatment, $\pi_k = \pi_\ell$, with precisely the additional fact that exactly one was treated, $Z_k + Z_\ell = 1$. Given that one of these subjects was treated, the chance that it was k and not ℓ would be

$$\Pr(Z_k = 1,\ Z_\ell = 0 \,|\, r_{Tk},\ r_{Ck},\ \mathbf{x}_k,\ u_k,\ r_{T\ell},\ r_{C\ell},\ \mathbf{x}_\ell,\ u_\ell,\ Z_k + Z_\ell = 1)$$

$$= \frac{\Pr(Z_k = 1,\ Z_\ell = 0 \,|\, r_{Tk},\ r_{Ck},\ \mathbf{x}_k,\ u_k,\ r_{T\ell},\ r_{C\ell},\ \mathbf{x}_\ell,\ u_\ell)}{\Pr(Z_k + Z_\ell = 1 \,|\, r_{Tk},\ r_{Ck},\ \mathbf{x}_k,\ u_k,\ r_{T\ell},\ r_{C\ell},\ \mathbf{x}_\ell,\ u_\ell)}$$

$$= \frac{\pi_\ell^{1+0}\ (1-\pi_\ell)^{(1-1)+(1-0)}}{\pi_\ell^{1+0}\ (1-\pi_\ell)^{(1-1)+(1-0)} + \pi_\ell^{0+1}\ (1-\pi_\ell)^{(1-0)+(1-1)}}$$

$$= \frac{\pi_\ell\ (1-\pi_\ell)}{\pi_\ell\ (1-\pi_\ell) + \pi_\ell\ (1-\pi_\ell)} = \frac{1}{2}$$

which, unlike (3.2), would be immensely useful. If we could find among the L subjects in the population a total of $2I$ different subjects matched in I pairs in just this way, then we would have reconstructed the distribution of treatment assignments $\mathbf{Z}$ in the paired, randomized experiment, and inference about the effects caused by treatments would be straightforward; indeed, the methods of Chap. 2 would be applicable.[3] Keep in mind that we are speaking here about mere imaginings, suppositions, hopes, aspirations, and attempts, as distinct from simple,

[2]As in Sect. 1.2.6, this definition of Z_k leaves open the possibility of using cohort size Z_k as an instrument for the actual size of the class, and indeed Angrist and Lavy [5] reported instrumental variable analyses.

[3]The argument just presented is the simplest version of a general argument developed in [81].

undeniable facts created by randomization in Chap. 2. In the words of the proverb: "If wishes were horses, beggars would ride."

Where do things stand? As has been seen, the ideal matching would pair individuals, k and ℓ, with different treatments, $Z_k + Z_\ell = 1$, but the same probability of treatment, $\pi_k = \pi_\ell$. If we could do this, we could reconstruct a randomized experiment from observational data; however, we cannot do this. The wise attempt to find "natural experiments" is the attempt to move towards this ideal. That is what you do with ideals: you attempt to move towards them. No doubt, many if not most "natural experiments" fail to produce this ideal pairing [10, 112]; moreover, even if one "natural experiment" once succeeded, there would be no way of knowing for sure that it had succeeded. A second attempt to approach the ideal is to find a context in which the covariates used to determine treatment assignment are measured covariates, $\mathbf{x}_k$, not unobserved covariates, u_ℓ, and then to match closely for the measured covariates, $\mathbf{x}_k$.[4] Matching for observed covariates is the topic of Part II of this book. Although we cannot see π_k and π_ℓ and so cannot see whether they are fairly close, perhaps there are things we can see that provide an incomplete and partial check on the closeness of π_k and π_ℓ. This leads to the devices of "quasi-experiments" [13, 15, 69, 119] mentioned in Sect. 1.2.4, such as multiple control groups [62, 84, 87], "multiple operationalism" or coherence [15, 47, 76, 90, 128], "control constructs" and known effects [68, 86, 87, 119, 133], and one or more pretreatment measures of what will become the response.[5] Even

[4]See, for instance, the argument made by David Gross and Nicholas Souleles, quoted in Sect. 4.2 from their paper [40]. The example in Chap. 8 attempts to use both approaches, although it is, perhaps, not entirely convincing in either attempt. In Chap. 8, the dose of chemotherapy received by patients with ovarian cancer is varied by comparing two types of providers of chemotherapy, gynecological oncologists (GO) and medical oncologists (MO), where the latter treat more intensively. The most important clinical covariates are measured, including clinical stage, tumor grade, histology, and comorbid conditions. Despite these two good attempts, it is far from clear that the GO-vs-MO assignment is haphazard, even after adjustments, and it is quite likely that the oncologist knew considerably more about the patient than is recorded in electronic form.

[5]Figure 1.2 is an example of quasi-experimental reasoning. Instead of comparing college attendance in two groups, say the children of deceased fathers in 1979–1981, when the benefit program was available, and the children of deceased fathers in 1982–1983, after the program was withdrawn, Fig. 1.2 follows Susan Dynarski's [26] study in comparing four groups. Figure 1.2 shows at most a small shift in college attendance by the group of children without deceased fathers, who were never eligible for the program. This is a small check, but a useful one: it suggests that the comparison of children of deceased fathers in 1979–1981 and 1982–1983, though totally confounded with time, may not be seriously biased by this confounding, because a substantial shift in college attendance by unaffected groups over this time is not evident. The quasi-experimental design in Fig. 1.2 is widely used and is known by a variety of not entirely helpful names, including "pretest–posttest nonequivalent control group design" and "difference-in-differences design." Kevin Volpp et al. [131] used this design in a medical context to study the possible effects on patient mortality of a reduction, to 80 h per week, of the maximum number of hours that a medical resident may work in a hospital. See [1, 7, 9, 14, 92] for discussion in general terms. The study by Garen Wintemute et al. [135], as discussed in Sect. 4.3, contains another example of quasi-experimental reasoning using a "control construct," that is, an outcome thought to be unaffected by the treatment. They were

when it "seems reasonably plausible that π_k and π_ℓ are fairly close," we will need to ask about the consequences of being close rather than equal, about the possibility they are somewhat less close than we had hoped, about the possibility that they are disappointingly far apart. Answers are provided by sensitivity analysis; see Sect. 3.4.

3.3 A Naïve Model: People Who Look Comparable Are Comparable

3.3.1 Assigning Treatments by Flipping Biased Coins Whose Unknown Biases Are Determined by Observed Covariates

The first of the two models in the title of this chapter asserts, rather naïvely, that people who look comparable in terms of measured covariates, $\mathbf{x}_k$, actually are comparable, and to a nontrivial extent, an element of luck determined every treatment assignment. Most observational studies report, perhaps briefly or perhaps at length, at least one analysis that tentatively assumes that the naïve model is true. Most of the controversy and skepticism associated with observational studies is a reference of one sort or another to this model's naïveté.

The naïve model restricts the representation (3.1) in two quite severe ways. First, it says that treatment assignment probabilities, $\pi_\ell = \Pr(Z_\ell = 1 \mid r_{T\ell}, r_{C\ell}, \mathbf{x}_\ell, u_\ell)$, that condition on $(r_{T\ell}, r_{C\ell}, \mathbf{x}_\ell, u_\ell)$ may depend upon the observed covariates $\mathbf{x}_\ell$ but not upon the potential responses $(r_{T\ell}, r_{C\ell})$ or the unobserved covariate u_ℓ. Second, the model says $0 < \pi_\ell < 1$ for all ℓ, so that every person ℓ has a chance of receiving either treatment, $Z_\ell = 1$, or control, $Z_\ell = 0$. Putting this together with (3.1), the naïve model says

$$\pi_\ell = \Pr(Z_\ell = 1 \mid r_{T\ell}, r_{C\ell}, \mathbf{x}_\ell, u_\ell) = \Pr(Z_\ell = 1 \mid \mathbf{x}_\ell) \tag{3.5}$$

and

$$0 < \pi_\ell < 1, \ \ell = 1, 2, \ldots, L, \tag{3.6}$$

interested in the possible effects of a change in gun control laws on violent crimes using guns. One would not expect a change in gun control laws to have a substantial effect on nonviolent, nongun crimes, and so the rate of such crimes is their control construct or unaffected outcome. See Sect. 4.3. An example with two control groups subject to different biases is discussed in Sect. 12.3.

with

$$\Pr(Z_1 = z_1, \ldots, Z_L = z_L | r_{T1}, r_{C1}, \mathbf{x}_1, u_1, \ldots, r_{TL}, r_{CL}, \mathbf{x}_L, u_L) \tag{3.7}$$

$$= \prod_{\ell=1}^{L} \pi_\ell^{z_\ell} (1 - \pi_\ell)^{1-z_\ell} . \tag{3.8}$$

The naïve model would be true if treatments Z_ℓ were assigned by independent flips of a fair coin: then $\pi_\ell = \frac{1}{2}$ for $\ell = 1, 2, \ldots, L$, and all of (3.5)–(3.8) is true. Indeed, the naïve model would be true if treatments Z_ℓ were assigned by independent flips of a group of biased coins, where the same biased coin is used whenever two people, k and ℓ, have the same observed covariate, $\mathbf{x}_k = \mathbf{x}_\ell$, and no coin has probability 1 or 0 of a head; then all of (3.5)–(3.8) are true. For (3.5)–(3.8) to be true, it is not necessary that the biases of these coins be known, but it is necessary that the bias depends on $\mathbf{x}_\ell$ alone. The first feature of the model (3.5)–(3.8) says treatment assignment Z is conditionally independent of (r_T, r_C, u) given $\mathbf{x}$, or in Phillip Dawid's [23] notation, $Z \perp\!\!\!\perp (r_T, r_C, u) \mid \mathbf{x}$. In [104], Donald Rubin and I gave the name "strongly ignorable treatment assignment given $\mathbf{x}$" to this naïve model (3.5)–(3.8).[6]

As will be seen in a moment, it would be very convenient if the naïve model (3.5)–(3.8) were true. If (3.5)–(3.8) were true, causal inference in an observational study would become a mechanical problem, so that if the sample size were large, and you followed certain steps, and you did not make mistakes in following the steps, then a correct causal inference would result; a computer could do it. So it would be convenient if (3.5)–(3.8) were true. To believe that something is true because it would be convenient if it were true is a fair definition of naïveté.

[6]Technically, strongly ignorable treatment assignment given $\mathbf{x}$ is $Z \perp\!\!\!\perp (r_T, r_C) \mid \mathbf{x}$ and $0 < \Pr(Z = 1 \mid \mathbf{x}) < 1$ for all $\mathbf{x}$, or in a single expression, $0 < \Pr(Z = 1 \mid r_T, r_C, \mathbf{x}) = \Pr(Z = 1 \mid \mathbf{x}) < 1$ for all $\mathbf{x}$; that is, no explicit reference is made to an unobserved covariate u. I introduce u here, rather than in Sect. 3.4 where it is essential, to simplify the transition between Sects. 3.3 and 3.4.

The remainder of this footnote is a slightly technical discussion of the formal relationship between (3.5) and the condition as given in [104]; it will interest only the insanely curious. In a straightforward way [23, Lemma 3], condition (3.5) implies $Z \perp\!\!\!\perp (r_T, r_C) \mid \mathbf{x}$ and $0 < \Pr(Z = 1 \mid \mathbf{x}) < 1$ for all $\mathbf{x}$. Also, in a straightforward way [23, Lemma 3], if (3.5) is true then (i) $Z \perp\!\!\!\perp (r_T, r_C) \mid (\mathbf{x}, u)$, (ii) $0 < \Pr(Z = 1 \mid \mathbf{x}) < 1$ for all $\mathbf{x}$, and (iii) $Z \perp\!\!\!\perp u \mid \mathbf{x}$ are true. Moreover, (ii) and (iii) together imply (iv) $0 < \Pr(Z = 1 \mid \mathbf{x}, u) < 1$ for all $(\mathbf{x}, u)$. In words, condition (3.5) implies both strong ignorability given $\mathbf{x}$ and also strong ignorability given $(\mathbf{x}, u)$. Now, conditions (i) and (iv), without condition (iii), are the key elements of the sensitivity model in Sect. 3.4: they say that treatment assignment would have been strongly ignorable if u had been measured and included with $\mathbf{x}$; so the failure to measure u is the source of our problems. Moreover, if condition (iii) were true in addition to (i) and (iv), then (3.5) and strong ignorability given $\mathbf{x}$ follow. In brief, strong ignorability given $\mathbf{x}$ is implied by the addition of condition (iii) to strong ignorability given $(\mathbf{x}, u)$. In words, one can reasonably think of the naive model Sect. 3.3 as the sensitivity model of Sect. 3.4 together with the irrelevance of u as expressed by $Z \perp\!\!\!\perp u \mid \mathbf{x}$ in condition (iii).

The Ideal Match and the Naïve Model

If the naïve model were true, then the ideal match of Sect. 3.2 could be produced simply by matching treated and control subjects with the same value of the observed covariate $\mathbf{x}$. This follows immediately from (3.5)–(3.8) and the discussion in Sect. 3.2. If the naïve model were true, one could reconstruct the distribution of treatment assignments $\mathbf{Z}$ in a randomized paired experiment from observational data simply by matching for the observed covariates, $\mathbf{x}$. Conventional methods of statistical analysis—the paired t-test, Wilcoxon's signed rank test, m-estimates, etc.—would work if (3.5)–(3.8) were true and are quite likely to fail if (3.5)–(3.8) is false in a nontrivial way.

What Is the Propensity Score?

So far, the talk has been of matching exactly for $\mathbf{x}$ as if that were easy to do. If $\mathbf{x}$ contains many covariates, then matching treated and control subjects with the same or similar $\mathbf{x}$ will be difficult if not impossible. Imagine, for instance, that $\mathbf{x}$ contains 20 covariates. Then any one subject, ℓ, might be above or below the median on the first covariate (two possibilities), above or below the median on the second covariate (two more possibilities, making $2 \times 2 = 4$ possibilities so far), and so on. Just in terms of being above or below the median on each covariate, there are $2 \times 2 \times \cdots \times 2 = 2^{20} = 1,048,576$ possibilities, or more than a million possibilities. With thousands of subjects, it will often be difficult to find two subjects who even match in the limited way of being on the same side of the median for all 20 covariates. This turns out to be less of a problem than it might seem at first, because of a device called the "propensity score" [104].

In a sample of L subjects from an infinite population, the propensity score [104] is the conditional probability of treatment $Z = 1$ given the observed covariates $\mathbf{x}$, or $e(\mathbf{x}) = \Pr(Z = 1 \mid \mathbf{x})$. Several aspects of this definition deserve immediate emphasis. The propensity score is defined in terms of the observed covariates, $\mathbf{x}$, whether or not the naïve model (3.5)–(3.8) is true, that is, whether or not the "ideal match" of Sect. 3.2 can be produced by matching on the observed covariates $\mathbf{x}$. The form of (3.5)–(3.8) correctly suggests that the propensity score $e(\mathbf{x}) = \Pr(Z = 1 \mid \mathbf{x})$ will be more useful when (3.5)–(3.8) is true; however, the propensity score is defined in terms of the observable quantities, Z and $\mathbf{x}$, whether or not (3.5)–(3.8) is true. In a randomized experiment, the propensity score is known because of random assignment. In contrast, in an observational study the propensity score is typically unknown; however, because $e(\mathbf{x}) = \Pr(Z = 1 \mid \mathbf{x})$ is defined in terms of observable quantities, namely treatment assignment Z and observed covariates, $\mathbf{x}$, it is straightforward to estimate the propensity score in an observational study. In a randomized experiment, π_ℓ in (3.1) is known and in an observational study π_ℓ is unknown, but because π_ℓ depends upon the unobservable (r_T, r_C, u) as well as $\mathbf{x}$, it is not possible to estimate π_ℓ in an observational study. Of course, if the naïve

model (3.5)–(3.8) were true, then $\pi_\ell = e(\mathbf{x}_\ell)$; indeed, this is virtually the defining feature of that model.

3.3.2 Balancing Property of the Propensity Score

The propensity score has several useful properties. The first property, the balancing property, is always true, whether or not the naïve model (3.5)–(3.8) is true. The balancing property [104] says that treated ($Z = 1$) and control ($Z = 0$) subjects with the same propensity score $e(\mathbf{x})$ have the same distribution of the observed covariates, $\mathbf{x}$,

$$\Pr\{\mathbf{x} \mid Z = 1,\ e(\mathbf{x})\} = \Pr\{\mathbf{x} \mid Z = 0,\ e(\mathbf{x})\} \tag{3.9}$$

or equivalently

$$Z \perp\!\!\!\perp \mathbf{x} \mid e(\mathbf{x}) \ , \tag{3.10}$$

so treatment Z and observed covariates $\mathbf{x}$ are conditionally independent given the propensity score.[7]

Because of the balancing property (3.10), if you pair two people, k and ℓ, one of whom is treated, $Z_k + Z_\ell = 1$, so that they have the same value of the propensity score, $e(\mathbf{x}_k) = e(\mathbf{x}_\ell)$, then they may have different values of the observed covariate, $\mathbf{x}_k \neq \mathbf{x}_\ell$, but in this pair, the specific values of the observed covariate $(\mathbf{x}_k, \mathbf{x}_\ell)$ will be unrelated to the treatment assignment (Z_k, Z_ℓ). If you form many pairs in this way, then the distribution of the observed covariates $\mathbf{x}$ will look about the same in the treated ($Z = 1$) and control ($Z = 0$) groups, even though individuals in matched pairs will typically have different values of $\mathbf{x}$. Although it is difficult to match on 20 covariates at once, it is easy to match on one covariate, the propensity score $e(\mathbf{x})$, and matching on $e(\mathbf{x})$ will tend to balance all 20 covariates.

In the randomized NSW experiment in Chap. 2, Table 2.1 displayed balance on observed covariates, $\mathbf{x}$. Randomization alone tends to balance observed covariates, $\mathbf{x}$, and Table 2.1 combined randomization and matching. In an observational study, it is often possible to match on an estimate $\widehat{e}(\mathbf{x})$ of the propensity score $e(\mathbf{x})$ and to

[7]The proof is easy [104]. Recall the following basic facts about conditional expectations: if A, B, and C are random variables, then $E(A) = E\{E(A \mid B)\}$ and $E(A \mid C) = E\{E(A \mid B, C) \mid C\}$. From the definition of conditional independence, to prove (3.10), it suffices to prove $\Pr\{Z = 1 \mid \mathbf{x}, e(\mathbf{x})\} = \Pr\{Z = 1 \mid e(\mathbf{x})\}$. Because $e(\mathbf{x})$ is a function of $\mathbf{x}$, conditioning on $\mathbf{x}$ fixes $e(\mathbf{x})$, so $\Pr\{Z = 1 \mid \mathbf{x}, e(\mathbf{x})\} = \Pr(Z = 1 \mid \mathbf{x}) = e(\mathbf{x})$. Because $Z = 0$ or $Z = 1$, for any random variable D, we have $\Pr(Z = 1 \mid D) = E(Z \mid D)$. With $A = Z$, $B = \mathbf{x}$, $C = e(\mathbf{x})$, we have $\Pr\{Z = 1 \mid e(\mathbf{x})\} = E\{Z \mid e(\mathbf{x})\} = E\{E\{Z \mid \mathbf{x}, e(\mathbf{x})\} \mid e(\mathbf{x})\} = E\{\Pr\{Z = 1 \mid \mathbf{x}, e(\mathbf{x})\} \mid e(\mathbf{x})\} = E\{e(\mathbf{x}) \mid e(\mathbf{x})\} = e(\mathbf{x}) = \Pr(Z = 1 \mid \mathbf{x})$, which is what we needed to prove.

produce balance on observed covariates $\mathbf{x}$ similar to the balance in Table 2.1; see, for example, Chap. 8.

Randomization is a much more powerful tool for balancing covariates than matching on an estimate of the propensity score. The difference is that the propensity score balances the observed covariates, $\mathbf{x}$, whereas randomization balances observed covariates, unobserved covariates, and potential responses, $(r_T, r_C, \mathbf{x}, u)$. This difference is not something you can see: randomization makes a promise about something you cannot see. Stated formally, if you match on $e(\mathbf{x})$, then $Z \perp\!\!\!\perp \mathbf{x} \mid e(\mathbf{x})$, but if you assign treatments by flipping fair coins, then $Z \not\perp\!\!\!\perp (r_T, r_C, \mathbf{x}, u)$. Randomization provides a basis for believing that an unobserved covariate u is balanced, but matching on propensity scores provides no basis for believing that.

Part II of this book discusses the practical aspects of matching, and an estimate $\widehat{e}(\mathbf{x})$ of the propensity score $e(\mathbf{x})$ is one of the tools used; see Sect. 9.2. [8] In common practice, the estimate of the propensity score, $\widehat{e}(\mathbf{x})$, is based on a model, such as a logit model, relating treatment assignment, Z, and observed covariates, $\mathbf{x}$. The logit model may include interactions, polynomials, and transformations of the covariates in $\mathbf{x}$, so it need not be a linear logit model narrowly conceived. Logit models for $\widehat{e}(\mathbf{x})$ are used for convenience; other methods can and have been used instead; , e.g., [34, 67, 137, 141]. After matching, balance on observed covariates is checked by comparing the distributions of observed covariates in treated and control groups, as in Tables 2.1 and 8.1; for general discussion, see Sect. 10.1. Because (3.10) is known to be true for the true propensity score, $e(\mathbf{x})$, a check on covariate balance, as in Table 8.1, is a diagnostic check of the model that produced the estimated propensity score, $\widehat{e}(\mathbf{x})$, and may lead to a revision of that model.

In most but not all cases, when you substitute an estimate, say $\widehat{e}(\mathbf{x})$, for the unknown true parameter, $e(\mathbf{x})$, the estimate performs somewhat less well than the true parameter would perform. In fact, this is not true for propensity scores: estimated scores, $\widehat{e}(\mathbf{x})$, tend to work slightly better than true scores, $e(\mathbf{x})$. Estimated scores tend to slightly overfit, producing slightly better than chance balance on observed covariates, $\mathbf{x}$, in the data set used to construct the score. Estimated scores produce "too much covariate balance," but because propensity scores are used to balance covariates, "too much covariate balance" is just fine.[9]

[8] Part II makes use of a stronger version of the balancing property than stated in (3.10). Specifically, if you match on $e(\mathbf{x})$ and any other aspect of $\mathbf{x}$, say $h(\mathbf{x})$, you still balance $\mathbf{x}$, that is, $Z \perp\!\!\!\perp \mathbf{x} \mid \{e(\mathbf{x}), h(\mathbf{x})\}$ for any $h(\mathbf{x})$. See [104] for the small adjustments required in the proof.

[9] Theoretical arguments also show that estimated propensity scores can have better properties than true propensity scores; see [81, 85].

3.3.3 Propensity Scores and Ignorable Treatment Assignment

The balancing property of propensity scores (3.10) is always true, but a second property of propensity scores would follow if the naïve model (3.5)–(3.8) were true. Recall that if (3.5)–(3.8) were true, then the "ideal match" of Sect. 3.2 could be produced simply by matching for the observed covariate $\mathbf{x}$. Recall also that it may be difficult to match closely for every one of the many covariates in $\mathbf{x}$, but it is easy to match on one variable, the propensity score, $e(\mathbf{x})$, and doing that balances all of $\mathbf{x}$. The second property closes this loop. It says: if (3.5)–(3.8) were true, then the "ideal match" of Sect. 3.2 could be produced by matching on the propensity score, $e(\mathbf{x})$, alone [104]. The proof is very short: the "ideal match" is to match on $\pi_\ell = \Pr(Z_\ell = 1 \mid r_{T\ell}, r_{C\ell}, \mathbf{x}_\ell, u_\ell)$, but if the naïve model (3.5)–(3.8) were true, then $\pi_\ell = e(\mathbf{x}_\ell)$, so matching on the propensity score is matching on π_ℓ. In words, if it suffices to match for the observed covariates, $\mathbf{x}$, then it suffices to match for the propensity score, $e(\mathbf{x})$.

The issue just discussed concerns a second property of propensity scores that holds when the naïve model (3.5)–(3.8) is true. This second property may be expressed in a different way. The naïve model (3.5)–(3.8) assumes

$$Z \perp\!\!\!\perp (r_T, r_C, u) \mid \mathbf{x}, \tag{3.11}$$

and the second property says that (3.11) implies

$$Z \perp\!\!\!\perp (r_T, r_C, u) \mid e(\mathbf{x}) \tag{3.12}$$

so the single variable $e(\mathbf{x})$ may be used in place of the many observed covariates in $\mathbf{x}$.[10]

3.3.4 Summary: Separating Two Tasks, One Mechanical, the Other Scientific

If the naïve model (3.5)–(3.8) were true, then the distribution of treatment assignments $\mathbf{Z}$ in a paired randomized experiment could be produced in an observational study by simply matching on observed covariates, $\mathbf{x}$. That might be a challenge if $\mathbf{x}$ contains many observed covariates, but in fact, matching on one covariate, the propensity score, $e(\mathbf{x})$, will balance all of the observed covariates, $\mathbf{x}$, and if

[10] In technical terms, if treatment assignment is strongly ignorable given $\mathbf{x}$, then it is strongly ignorable given $e(\mathbf{x})$. The proof of this version is also short [104]. In footnote 7, we saw that $\Pr\{Z = 1 \mid e(\mathbf{x})\} = \Pr(Z = 1 \mid \mathbf{x})$. Condition (3.11) says $\Pr(Z = 1 \mid r_T, r_C, \mathbf{x}, u) = \Pr(Z = 1 \mid \mathbf{x})$. Together they say $\Pr(Z = 1 \mid r_T, r_C, \mathbf{x}, u) = \Pr\{Z = 1 \mid e(\mathbf{x})\}$, which implies (3.12).

(3.5)–(3.8) were true, this match would also produce the distribution of treatment assignments **Z** in a paired randomized experiment. Better still, it is straightforward to check whether matching has balanced the observed covariates; one simply checks whether matched treated and control groups are similar in terms of **x**.

The naïve model (3.5)–(3.8) is important, but not because it is plausible; it is not plausible. The controversy and skepticism that almost invariably attends an observational study almost invariably refers back to this model's naïveté. Rather, the naïve model (3.5)–(3.8) is important because it cleanly divides inference in observational studies into two separable tasks. One is a fairly mechanical task that typically can be completed successfully and can be seen to have been completed successfully, before the second task is engaged. This first task can take the form of matching treated and control subjects so that observed covariates are seen to be balanced. The first task is to compare treated and control subjects who look comparable prior to treatment. The second task engages the concern that people who look comparable may not be comparable. If people were not randomly assigned to treatments, then perhaps there are reasons they received the treatments they did, but those reasons are not visible to us because the observed covariates **x** provide an incomplete picture of the situation prior to treatment. The second task is not a mechanical but rather a scientific task, one that can be controversial and difficult to bring to a rapid and definitive closure; this task is, therefore, more challenging, and hence more interesting. The clever opportunities of natural experiments, the subtle devices of quasi-experiments, the technical tools of sensitivity analysis that extract information from nonidentified models — each of these is an attempt to engage the second task.

3.4 Sensitivity Analysis: People Who Look Comparable May Differ

3.4.1 What Is Sensitivity Analysis?

If the naïve model (3.5)–(3.8) were true, the distribution of treatment assignments **Z** in a randomized paired experiment could be reconstructed by matching for the observed covariate, **x**. It is common for a critic to argue that, in a particular study, the naïve model may be false. Indeed, it may be false. Typically, the critic accepts that the investigators matched for the observed covariates, **x**, so treated and control subjects are seen to be comparable in terms of **x**, but the critic points out that the investigators did not measure a specific covariate u, did not match for u, and so are in no position to assert that treated and control groups are comparable in terms of u. This criticism could be dismissed in a randomized experiment—randomization does tend to balance unobserved covariates—but the criticism cannot be dismissed in an observational study. This difference in the unobserved covariate u, the critic continues, is the real reason outcomes differ in the treated and control

groups: it is not an effect caused by the treatment, but rather a failure on the part of the investigators to measure and control imbalances in u. Although not strictly necessary, the critic is usually aided by an air of superiority: "This would never happen in my laboratory."

It is important to recognize at the outset that our critic may be, but need not be, on the side of the angels. The tobacco industry and its (sometimes distinguished) consultants criticized, in precisely this way, observational studies linking smoking with lung cancer [125]. In this instance, the criticism was wrong. Investigators and their critics stand on level ground [11].

It is difficult if not impossible to give form to arguments of this sort until one has a way of speaking about the degree to which the naïve model is false. In an observational study, one could never assert with warranted conviction that the naïve model is precisely true. Trivially small deviations from the naïve model will have a trivially small impact on the study's conclusions. Sufficiently large deviations from the naïve model will overturn the results of any study. Because these two facts are always true, they quickly exhaust their usefulness. Therefore, the magnitude of the deviation is all-important. The sensitivity of an observational study to bias from an unmeasured covariate u is the magnitude of the departure from the naïve model that would need to be present to materially alter the study's conclusions.[11]

The first sensitivity analysis in an observational study concerned smoking and lung cancer. In 1959, Jerry Cornfield and his colleagues [20] asked about the magnitude of the bias from an unobserved covariate u needed to alter the conclusion from observational studies that heavy smoking causes lung cancer. They concluded that the magnitude of the bias would need to be enormous.

3.4.2 The Sensitivity Analysis Model: Quantitative Deviation from Random Assignment

The naïve model (3.5)–(3.8) said that two people, k and ℓ, with the same observed covariates, $\mathbf{x}_k = \mathbf{x}_\ell$, have the same probability of treatment given $(r_T, r_C, \mathbf{x}, u)$, i.e., $\pi_k = \pi_\ell$, where $\pi_k = \Pr(Z_k = 1 \mid r_{Tk}, r_{Ck}, \mathbf{x}_k, u_k)$ and $\pi_\ell = \Pr(Z_\ell = 1 \mid r_{T\ell}, r_{C\ell}, \mathbf{x}_\ell, u_\ell)$. The sensitivity analysis model speaks about the same probabilities in (3.1), saying that the naïve model (3.5)–(3.8) may be false, but to an extent controlled by a parameter, $\Gamma \geq 1$. Specifically, it says that

[11] In general, a sensitivity analysis asks how the conclusion of an argument dependent upon assumptions would change if the assumptions were relaxed. The term is sometimes misused to refer to performing several parallel statistical analyses without regard to the assumptions upon which they depend. If several statistical analyses all depend upon the same assumption—for instance, the naïve model (3.5)—then performing several such analyses provides no insight into consequences of the failure of that assumption.

two people, k and ℓ, with the same observed covariates, $\mathbf{x}_k = \mathbf{x}_\ell$, have odds[12] of treatment, $\pi_k / (1 - \pi_k)$ and $\pi_\ell / (1 - \pi_\ell)$, that differ by at most a multiplier of Γ; that is, in (3.1),

$$\frac{1}{\Gamma} \leq \frac{\pi_k / (1 - \pi_k)}{\pi_\ell / (1 - \pi_\ell)} \leq \Gamma \text{ whenever } \mathbf{x}_k = \mathbf{x}_\ell \,. \tag{3.13}$$

If $\Gamma = 1$ in (3.13), then $\pi_k = \pi_\ell$, so (3.5)–(3.8) is true; that is, $\Gamma = 1$ corresponds with the naïve model. In Sect. 3.1, expression (3.1) was seen to be a representation and not a model—something that is always true for suitably defined u_ℓ—but that representation took $\pi_\ell = 0$ or $\pi_\ell = 1$, which implies $\Gamma = \infty$ in (3.13). In other words, numeric values of Γ between $\Gamma = 1$ and $\Gamma = \infty$ define a spectrum that begins with the naïve model (3.5)–(3.8) and ends with something that is hollow in the sense that it is always true, namely (3.1). The hollow statement that is always true, namely (3.1), is the statement that "association does not imply causation," that is, a sufficiently large departure from the naïve model can explain away as noncausal any observed association.

If $\Gamma = 2$, and if you, k , and I, ℓ, look the same, in the sense that we have the same observed covariates, $\mathbf{x}_k = \mathbf{x}_\ell$, then you might be twice as likely as I to receive the treatment because we differ in ways that have not been measured. For instance, if your $\pi_k = 2/3$ and my $\pi_\ell = 1/2$, then your odds of treatment rather than control are $\pi_k / (1 - \pi_k) = 2$ or 2-to-1, whereas my odds of treatment rather than control are $\pi_\ell / (1 - \pi_\ell) = 1$ or 1-to-1, and you are twice as likely as I to receive treatment, $\{\pi_k / (1 - \pi_k)\} / \{\pi_\ell / (1 - \pi_\ell)\} = 2$ in (3.13).[13]

[12]Odds are an alternative way of expressing probabilities. Probabilities and odds carry the same information in different forms. A probability of $\pi_k = 2/3$ is an odds of $\pi_k / (1 - \pi_k) = 2$ or 2-to-1. Gamblers prefer odds to probabilities because odds express the chance of an event in terms of fair betting odds, the price of a fair bet. It is easy to move from probability π_k to odds $\omega_k = \pi_k / (1 - \pi_k)$ and back again from odds ω_k to probability $\pi_k = \omega_k / (1 + \omega_k)$.

[13]Implicitly, the critic is saying that the failure to measure u is the source of the problem, or that (3.5) would be true with $(\mathbf{x}, u)$ in place of $\mathbf{x}$, but is untrue with $\mathbf{x}$ alone. That is, the critic is saying $\pi_\ell = \Pr(Z_\ell = 1 \mid r_{T\ell}, r_{C\ell}, \mathbf{x}_\ell, u_\ell) = \Pr(Z_\ell = 1 \mid \mathbf{x}_\ell, u_\ell)$. As in Sect. 3.1, because of the delicate nature of unobserved variables, this is a manner of speaking rather than a tangible distinction. If the formalities are understood to refer to $\pi_\ell = \Pr(Z_\ell = 1 \mid r_{T\ell}, r_{C\ell}, \mathbf{x}_\ell, u_\ell)$, then it is not necessary to insist that $\pi_\ell = \Pr(Z_\ell = 1 \mid \mathbf{x}_\ell, u_\ell)$. Conversely, there is always a scalar unobserved covariate u with $0 \leq u \leq 1$ such that $Z \perp\!\!\!\perp (r_T, r_C) \mid (\mathbf{x}, u)$, namely the unobserved covariate $u = \zeta$, where $\zeta = \Pr(Z = 1 | r_T, r_C, \mathbf{x})$; for proof, see [102, §6.3]. Consistent with terminology introduced by Frangakis and Rubin [33], we might call this unobserved covariate, $u = \zeta$, the "principal unobserved covariate." In effect, the bounded, scalar principal unobserved covariate, $u = \zeta$, is the only unobserved covariate that matters, and it always exists. A finite Γ changes the principal unobserved covariate, $u = \zeta$, from a representation to a model, because a finite Γ implies $0 < \zeta < 1$ rather than $0 \leq \zeta \leq 1$.

3.4.3 Sensitivity Analysis Model When Pairs Are Matched for Observed Covariates

The sensitivity analysis model (3.13) is quite general in its applicability ([93, Chapter 4] and [101]), but here its implications for matched pairs are developed [83]. Suppose that two subjects, k and ℓ, with the same observed covariates, $\mathbf{x}_k = \mathbf{x}_\ell$, are paired, with precisely the additional fact that one of them is treated and the other control, $Z_k + Z_\ell = 1$. Then in the representation (3.1), the chance that k is treated and ℓ is control is $\Pr(Z_k = 1,\ Z_\ell = 0 \,|\, r_{Tk},\ r_{Ck},\ \mathbf{x}_k,\ u_k,\ r_{T\ell},\ r_{C\ell},\ \mathbf{x}_\ell,\ u_\ell,\ Z_k + Z_\ell = 1)$

$$= \frac{\pi_k(1-\pi_\ell)}{\pi_k(1-\pi_\ell) + \pi_\ell(1-\pi_k)}. \tag{3.14}$$

If in addition the sensitivity model (3.13) were true in (3.1), then simple algebra yields

$$\frac{1}{1+\Gamma} \le \frac{\pi_k(1-\pi_\ell)}{\pi_k(1-\pi_\ell) + \pi_\ell(1-\pi_k)} \le \frac{\Gamma}{1+\Gamma}. \tag{3.15}$$

In words, the condition (3.13) becomes a new condition (3.15) on paired individuals where one is treated and the other control, $Z_k + Z_\ell = 1$. If $\Gamma = 1$, then all three terms in (3.15) equal $\frac{1}{2}$, as in the randomized experiment in Chap. 2. As $\Gamma \to \infty$, the lower bound in (3.13) tends to zero and the upper bound tends to one.

Instead of pairing just two individuals, k and ℓ, suppose we pair $2I$ distinct individuals of the L individuals in the population in just this way, insisting that within each pair the two subjects have the same observed covariates and different treatments. Renumber these paired subjects into I pairs of two subjects, $i = 1, 2, \ldots, I$, $j = 1, 2$, so $\mathbf{x}_{i1} = \mathbf{x}_{i2}$, $Z_{i1} = 1 - Z_{i2}$ in each of the I pairs.[14] If (3.1) and (3.13) are true, then the distribution of treatment assignments in the I pairs satisfies

$$Z_{i1},\ i = 1, \ldots, I \ \text{ are mutually independent,} \tag{3.16}$$

$$Z_{i2} = 1 - Z_{i1},\ i = 1, \ldots, I, \tag{3.17}$$

[14] In a fussy technical sense, the numbering of pairs and people within pairs is supposed to convey nothing about these people, except that they were eligible to be paired, that is, they have the same observed covariates, different treatments, with $2I$ distinct people. Information about people is supposed to be recorded in variables that describe them, such as $Z, \mathbf{x}, u, r_T, r_C$, not in their position in the data set. You cannot put your brother-in-law in the last pair just because of that remark he made last Thanksgiving; you have to code him in an explicit brother-in-law variable. Obviously, it is easy to make up subscripts that meet this fussy requirement: number the pairs at random, then number the people in a pair at random. The fussy technical point is that, in going from the L people in (3.1) to the $2I$ paired people, no information has been added and tucked away into the subject numbers—the criteria for pairs are precisely $\mathbf{x}_{i1} = \mathbf{x}_{i2}$, $Z_{i1} + Z_{i2} = 1$ with $2I$ distinct individuals.

$$\frac{1}{1+\Gamma} \leq \frac{\pi_k(1-\pi_\ell)}{\pi_k(1-\pi_\ell)+\pi_\ell(1-\pi_k)} \leq \frac{\Gamma}{1+\Gamma},\ i = 1, \ldots, I. \tag{3.18}$$

This is very similar in form to the distribution of treatment assignments in a randomized paired experiment in Chap. 2, except that in the experiment (3.14) was $\frac{1}{2}$ for $i = 1, \ldots, I$, whereas in (3.16)–(3.18) the treatment assignment probabilities may vary from pair to pair, are unknown, but are bounded by $1/(1+\Gamma)$ and $\Gamma/(1+\Gamma)$. If $\Gamma = 1.0001$, then (3.14) would differ trivially from a randomized paired experiment, but as $\Gamma \to \infty$ the difference can become arbitrarily large.

Suppose that we had calculated a P-value or a point estimate or confidence interval from a paired observational study matched for observed covariates $\mathbf{x}$, by simply applying conventional statistical methods, that is, the methods in Chap. 2 for a randomized paired experiment. Those inferences would have their usual properties if the naïve model (3.5)–(3.8) were true, that is, if $\Gamma = 1$. How might those inferences change if Γ were some specific number larger than 1, indicating some bias due to failure to control for u? Using (3.16)–(3.18) and a few calculations, we can often deduce the range of possible P-values or point estimates or confidence intervals for a specified Γ. Consider, for instance, the P-value for testing the null hypothesis of no treatment effect. If the naïve model, $\Gamma = 1$, led to a P-value of, say, 0.001, and if $\Gamma = 2$ yields a range of possible P-values from 0.0001 to 0.02, then a bias of magnitude $\Gamma = 2$ creates greater uncertainty but does not alter the qualitative conclusion that the null hypothesis of no effect has been rejected at level $\alpha = 0.05$.

Every study is sensitive to sufficiently large biases. There is always a value of Γ such that, for that value and larger values of Γ, the interval of possible P-values includes small values, perhaps 0.0001, and large values, perhaps 0.1. A sensitivity analysis simply displays how the inference changes with Γ. For smoking and lung cancer, the bias would have to be enormous, $\Gamma = 6$; see [93, Chapter 4]. The question answered by a sensitivity analysis is: How large does Γ have to be before one must concede that the critic's criticism might be correct? Is the critic's criticism outlandish? If so, precisely how outlandish?

It is time to consider an example.

3.5 Welding Fumes and DNA Damage

3.5.1 Sensitivity Analysis When Testing the Hypothesis of No Treatment Effect

The fumes produced by electric welding contain chromium and nickel and have been judged genotoxic in laboratory tests [50]. Werfel et al. [134] looked for evidence of DNA damage in humans by comparing 39 male welders to 39 male controls matched for age and smoking habits. Table 3.1 displays the comparability of the

Table 3.1 Covariate balance in 39 matched welder–control pairs

		Welders	Controls
Male		100%	100%
Smokers		69%	69%
Age	Mean	39	39
	Minimum	23	23
	Lower Quartile	34	32
	Median	38	36
	Upper Quartile	46	46
	Maximum	56	59

Covariates are gender, Smoking, and age

two groups with respect to the three covariates used in matching. Clearly, Table 3.1 is a rather limited demonstration of comparability.

Werfel et al. [134] presented several measures of genetic damage, including the measurement of DNA single strand breakage and DNA–protein cross-links using elution rates through polycarbonate filters with proteinase K. Broken strands are expected to pass through filters more quickly, at a higher rate. Figure 3.1 depicts the elution rates and their matched pair differences. The differences are mostly positive, with higher elution rates for welders, and the differences are fairly symmetric about their median, with longer tails than the Normal distribution.

Table 3.2 is the sensitivity analysis for the one-sided P-value using Wilcoxon's signed rank statistic to test the null hypothesis of no treatment effect against the alternative that exposure to welding fumes caused an increase in DNA damage. The first row, $\Gamma = 1$, is the usual randomization inference, which would be appropriate if the 78 men had been paired for age and smoking and randomly assigned to their careers as a welder or a nonwelder. In the first row, the range of possible P-values is a single number, 3.1×10^{-7}, because there would be no uncertainty about the distribution of treatment assignments, $\mathbf{Z}$, in a randomized experiment. The naïve model (3.5)–(3.8) would also lead to $\Gamma = 1$ and the single P-value in the first row of Table 3.2. If this had been a randomized experiment, there would have been strong evidence against the null hypothesis of no effect. However, it was not a randomized experiment. The P-value in the first row of Table 3.2 says that it is implausible that the difference seen in Fig. 3.1 is due to chance, the flip of a coin that assigned one man to treatment, another to control. The P-value in the first row of Table 3.2 does not speak to the critic's concern that the difference seen in Fig. 3.1 is neither due to chance nor due to an effect caused by welding, but reflects instead some way that the matched welders and controls are not comparable. A small P-value, here 3.1×10^{-7}, computed assuming either randomization or equivalently the naïve model (3.5)–(3.8) does nothing to address the critic's concern. It is, however, possible to speak to that concern.

The second row permits a substantial departure from random treatment assignment or (3.5)–(3.8). It says that two men of the same age and smoking status—the same $\mathbf{x}$—may not have the same chance of a career as a welder: one such man may be twice as likely as another to choose a career as a welder, $\Gamma = 2$, because

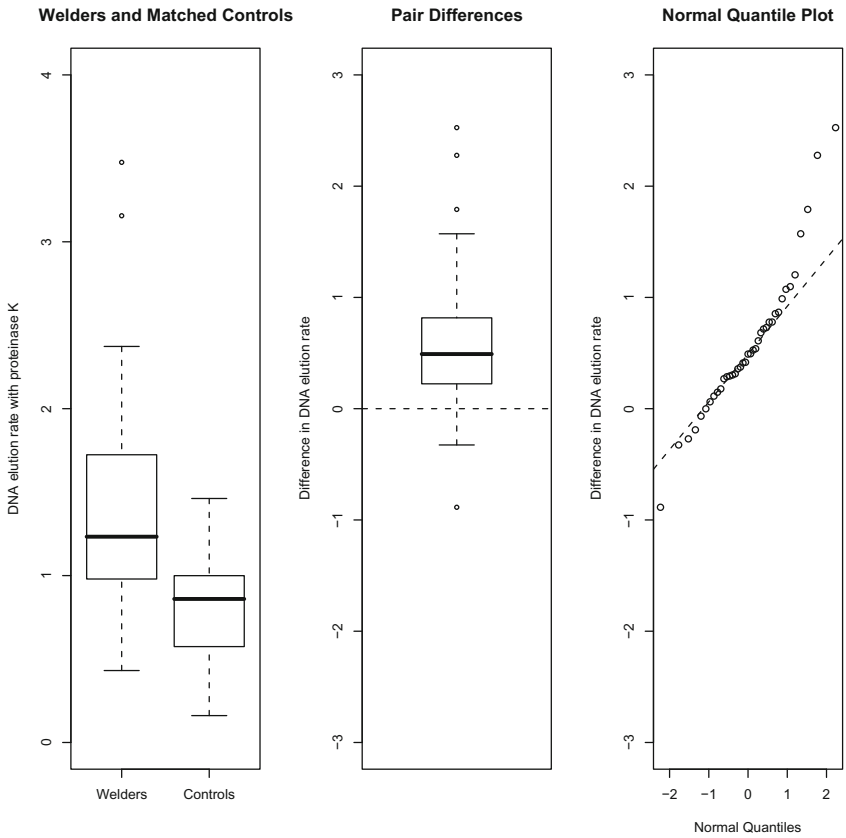

Fig. 3.1 DNA elution rates through polycarbonate filters with proteinase K for 39 male welders and 39 male controls matched for age and smoking. This assay is a measure of DNA single strand breakage and DNA–protein cross-links. In the boxplot of differences, there is a line at zero. In the Normal quantile plot, the line is fitted to the median and quartiles

they differ in terms of a covariate u that was not measured. This introduces a new source of uncertainty beyond chance. Using (3.16)–(3.18), we may determine every possible P-value that could be produced when $\Gamma = 2$, and it turns out that the smallest possible P-value is 3.4×10^{-12} and the largest possible P-value is 0.00064. Although a bias of magnitude $\Gamma = 2$ would introduce greater uncertainty, there would be no real doubt that the null hypothesis of no treatment effect is not plausible.

As seen in Table 3.2, all possible P-values are less than 0.05 for departures from randomization as large as $\Gamma = 4$. A bias of $\Gamma = 4$ is a very large departure from a randomized experiment. In a randomized experiment, each man in each pair has probability $\frac{1}{2}$ of receiving treatment. If $\Gamma = 4$, then in a matched pair, one man might have probability $\Gamma/(1+\Gamma) = 4/5$ of treatment and the other might

Table 3.2 Sensitivity analysis for the one-tailed P-value for testing the null hypothesis of no treatment effect on DNA elution rates with proteinase K in 39 pairs of a male welder and a male control matched for age and smoking

	$P_{\min}$	$P_{\max}$
Γ		
1	3.1×10^{-7}	3.1×10^{-7}
2	3.4×10^{-12}	0.00064
3	$<10^{-15}$	0.011
4	$<10^{-15}$	0.047
5	$<10^{-15}$	0.108

The table gives the lower (min) and upper (max) bounds on the one-sided P-value for departures from random assignment of various magnitudes, Γ. For $\Gamma = 1$, the two P-values are equal to each other and equal to the randomization P-value from Chap. 2. For $\Gamma > 1$, there is a range $[P_{\min}, P_{\max}]$ of possible P-values. This study is sensitive only to very large biases, for instance, $\Gamma = 5$, because at this point the range includes both small and large, significant and insignificant, P-values

have probability $1/(1+\Gamma) = 1/5$; however, even such a large departure from a randomized experiment is quite unlikely to produce the difference seen in Fig. 3.1.

By $\Gamma = 5$, the situation has changed. Now the range of possible P-values includes some that are much smaller than the conventional 0.05 level and others that are considerably higher, the range being from $< 10^{-15}$ to 0.108. A very large departure from random assignment of magnitude $\Gamma = 5$ could produce the difference seen in Fig. 3.1 even if welding has no effect on DNA elution rates.

3.5.2 Computations

The P-values in Table 3.2 are exact: in principle, they could be produced by direct enumeration analogous to those in Sect. 2.3.3, except that, now, different treatment assignments $\mathbf{Z}$ have different probabilities constrained by (3.16)–(3.18). The actual exact computations are done more efficiently; see Appendix.

Look back at (3.16)–(3.18) and consider how to set the probabilities to make T as large or as small as possible. In the absence of ties, the upper bound in Table 3.2 is obtained by comparing Wilcoxon's signed rank statistic to the distribution of a random variable, $\overline{\overline{T}}$, which is the sum of I independent random variables, $i = 1, 2, \ldots, I$, that take the value i with probability $\Gamma/(1+\Gamma)$ or the value 0 with probability $1/(1+\Gamma)$; see [83]. In parallel, the lower bound in Table 3.2 is obtained by comparing Wilcoxon's signed rank statistic to the distribution of a

random variable, $\overline{T}$, which is the sum of I independent random variables, $i = 1, 2, \ldots, I$, that take the value i with probability $1/(1+\Gamma)$ or the value 0 with probability $\Gamma/(1+\Gamma)$.

Although exact computation is quite feasible for moderate sample sizes, I, a large sample approximation is easier and typically adequate. In Table 3.2, Wilcoxon's signed rank statistic is $T = 715$ and there are no ties. In the absence of ties, for a specified Γ, the largest null distribution[15] of T subject to (3.16)–(3.18) has expectation

$$E\left(\overline{\overline{T}} \middle| \mathscr{F}, \mathscr{Z}\right) = \frac{\Gamma}{1+\Gamma} \cdot \frac{I(I+1)}{2} \tag{3.19}$$

and variance

$$\operatorname{var}\left(\overline{\overline{T}} \middle| \mathscr{F}, \mathscr{Z}\right) = \frac{\Gamma}{(1+\Gamma)^2} \frac{I(I+1)(2I+1)}{6}. \tag{3.20}$$

For $\Gamma = 1$, formulas (3.19) and (3.20) reduce to the formulas for randomization inference in Sect. 2.3.3, namely $E(T|\mathscr{F}, \mathscr{Z}) = I(I+1)/4$ and $\operatorname{var}(T|\mathscr{F}, \mathscr{Z}) = I(I+1)(2I+1)/24$. For $\Gamma = 3$, the expectation is

$$E\left(\overline{\overline{T}} \middle| \mathscr{F}, \mathscr{Z}\right) = \frac{3}{(1+3)} \cdot \frac{39(39+1)}{2} = 585 \tag{3.21}$$

in (3.19) and the variance is

$$\operatorname{var}\left(\overline{\overline{T}} \middle| \mathscr{F}, \mathscr{Z}\right) = \frac{3}{(1+3)^2} \frac{39(39+1)(2 \cdot 39+1)}{6} = 3851.25 \tag{3.22}$$

[15] What does it mean to speak of the "largest distribution"? One random variable, A, is said to be stochastically larger than another random variable, B, if $\Pr(A \geq k) \geq \Pr(B \geq k)$ for every k. That is, A is more likely than B to jump over a bar at height k, no matter how high, k, the bar is set. Because this must be true *for every* k, it is a rather special relationship between random variables. For instance, it might happen that $\Pr(A \geq -1.65) = 0.95$, $\Pr(A \geq 1.65) = 0.05$ while $\Pr(B \geq -1.65) = 0.80$, $\Pr(B \geq 1.65) = 0.20$, so neither A nor B is stochastically larger than the other. For instance, this is true if A is Normal with mean zero and standard deviation 1, and B is Normal with mean zero and standard deviation 2. The intuition is overwhelmingly strong that the largest null distribution of Wilcoxon's signed rank statistic is obtained by driving the chance of a positive difference up to its maximum, namely $\Gamma/(1+\Gamma)$ in (3.18), and this intuition turns out to be correct. It takes a small amount of effort in this case to show the distribution is actually stochastically largest; see [93, Chapter 4]. In other cases, the same result is available with somewhat more effort. In still other cases, one can speak of a "largest distribution" only asymptotically, that is, only in large samples; see [37, 101] or [93, Chapter 4]. This presents no problem in practice, because the asymptotic results are quite adequate and easy to use [93, Chapter 4]; however, it does make the theory of the paired case simpler than, say, the theory of matching with several controls matched to each treated subject. To see a parallel discussion of the paired case and the case of several controls, see [97]. For technical details, see [101].

in (3.20). For large I, the standardized deviate

$$\frac{T - E\left(\overline{\overline{T}}\middle|\, \mathscr{F}, \mathscr{Z}\right)}{\sqrt{\operatorname{var}\left(\overline{\overline{T}}\middle|\, \mathscr{F}, \mathscr{Z}\right)}} = \frac{715 - 585}{\sqrt{3851.25}} = 2.0948 \tag{3.23}$$

is compared to the standard Normal cumulative distribution, $\Phi(\cdot)$, to yield the approximate upper bound on the one-sided P-value, $1 - \Phi(2.0948) = 0.018$, which is close to the exact value of 0.011 in Table 3.2. The lower bound is obtained in a parallel manner, with

$$E\left(\overline{T}\middle|\, \mathscr{F}, \mathscr{Z}\right) = \frac{1}{1+\Gamma} \cdot \frac{I(I+1)}{2} \tag{3.24}$$

and $\operatorname{var}\left(\overline{T}\middle|\, \mathscr{F}, \mathscr{Z}\right) = \operatorname{var}\left(\overline{\overline{T}}\middle|\, \mathscr{F}, \mathscr{Z}\right)$ given again by (3.20).

Ties are a minor inconvenience. In the notation of Sect. 2.3.3, the expectations become

$$E\left(\overline{\overline{T}}\middle|\, \mathscr{F}, \mathscr{Z}\right) = \frac{\Gamma}{1+\Gamma} \sum_{i=1}^{I} s_i\, q_i\,, \tag{3.25}$$

$$E\left(\overline{T}\middle|\, \mathscr{F}, \mathscr{Z}\right) = \frac{1}{1+\Gamma} \sum_{i=1}^{I} s_i\, q_i\,, \tag{3.26}$$

while the variance becomes

$$\operatorname{var}\left(\overline{T}\middle|\, \mathscr{F}, \mathscr{Z}\right) = \operatorname{var}\left(\overline{\overline{T}}\middle|\, \mathscr{F}, \mathscr{Z}\right) = \frac{\Gamma}{(1+\Gamma)^2} \sum_{i=1}^{I} (s_i\, q_i)^2\,. \tag{3.27}$$

The remaining calculations are unchanged.

3.5.3 *Sensitivity Analysis for a Confidence Interval*

Table 3.3 is the sensitivity analysis for the one-sided 95% confidence interval for an additive, constant treatment effect discussed in Sect. 2.4.2. As in a randomized experiment, the hypothesis that $H_0 : r_{Tij} = r_{Cij} + \tau_0$ is tested by testing the null hypothesis of no treatment effect on the adjusted responses, $R_{ij} - \tau_0\, Z_{ij}$, or equivalently on the adjusted, treated-minus-control pair differences, $Y_i - \tau_0$. The one-sided 95% confidence interval is the set of values of τ_0 not rejected by a one-sided, 0.05 level test.

Table 3.3 Sensitivity analysis for the one-sided 95% confidence interval for a constant, additive treatment effect τ on DNA elution rates

Γ	1	2	3
95% Interval	$[0.37, \infty)$	$[0.21, \infty)$	$[0.094, \infty)$

As usual, the hypothesis of a constant effect $H_0 : \tau = \tau_0$ is tested by testing no effect on $Y_i - \tau_0$ for the given value of Γ. The one-sided 95% confidence interval is the set of values of τ_0 not rejected in the one-sided, 0.05 level test. As Γ increases, there is greater potential deviation from random treatment assignment in (3.13), and the confidence interval grows longer. For instance, a treatment effect of $\tau_0 = 0.30$ would be implausible in a randomized experiment, $\Gamma = 1$, but not in an observational study with $\Gamma = 2$

From Table 3.2, the hypothesis $H_0 : \tau = \tau_0$ for $\tau_0 = 0$ is barely rejected for $\Gamma = 4$ because the maximum possible one-sided P-value is 0.047. For $\Gamma = 3$, the maximum possible one-sided P-value is 0.04859 for $\tau_0 = .0935$ and is 0.05055 for $\tau_0 = .0936$, so after rounding to two significant digits, the one-sided 95% confidence interval is $[0.094, \infty)$.

3.5.4 Sensitivity Analysis for Point Estimates

For each value of $\Gamma \geq 1$, a sensitivity analysis replaces a single point estimate, say $\widehat{\tau}$, by an interval of point estimates, say $[\widehat{\tau}_{\min}, \widehat{\tau}_{\max}]$ that are the minimum and maximum point estimates for all distributions of treatment assignments satisfying (3.16)–(3.18). Unlike a test or a confidence interval, and like a point estimate, this interval $[\widehat{\tau}_{\min}, \widehat{\tau}_{\max}]$ does not reflect sampling uncertainty; however, it does reflect uncertainty introduced by departures from random treatment assignment in (3.13) or (3.16)–(3.18).

In a randomized experiment in Sect. 2.4.3, the Hodges–Lehmann point estimate of a constant, additive treatment effect, τ, was obtained by computing Wilcoxon's signed rank statistic T from $Y_i - \tau_0$ and solving for the estimate, $\widehat{\tau}$, as the value of τ_0 that brings T as close as possible to its null expectation, $I(I+1)/4$. It is not difficult to show[16] the interval of point estimates, $[\widehat{\tau}_{\min}, \widehat{\tau}_{\max}]$, is obtained by finding the value $\widehat{\tau}_{\min}$ such that when Wilcoxon's T is computed from $Y_i - \widehat{\tau}_{\min}$ it is as close as possible to (3.19), and the value $\widehat{\tau}_{\max}$ such that when T is computed from $Y_i - \widehat{\tau}_{\max}$ it is as close as possible to (3.24). For example, with $\Gamma = 2$, the maximum expectation in (3.19) is $\{\Gamma/(1+\Gamma)\} I(I+1)/2$ or $\{2/(1+2)\} 39(39+1)/2 = 520$. When Wilcoxon's T is computed from $Y_i - 0.35550001$ it is $T = 519$, but when it is computed from $Y_i - 0.35549999$ it is $T = 521$, so $\widehat{\tau}_{\min} = 0.3555$.

[16] See [88] or [93, Chapter 4].

Table 3.4 Sensitivity analysis for the Hodges–Lehmann (HL) point estimate for a constant, additive treatment effect τ on DNA elution rates

Γ	1	2	3
HL estimate	[0.51, 0.51]	[0.36, 0.69]	[0.27, 0.81]

Table 3.4 displays $[\widehat{\tau}_{\min}, \widehat{\tau}_{\max}]$ for three values of Γ. For $\Gamma = 1$, there is a single point estimate, $\widehat{\tau} = 0.51$, the value that would be obtained in a randomized experiment. A large departure from random assignment of magnitude $\Gamma = 3$ could reduce that by almost half, to $\widehat{\tau}_{\min} = 0.27$.

3.6 Amplification of a Sensitivity Analysis

3.6.1 What Is an Amplification?

An amplification of a sensitivity analysis [110] is a reinterpretation of a sensitivity analysis defined in terms of one parameter Γ. It is not a new sensitivity analysis; rather, it is an alternative description of the sensitivity analysis already performed in terms of $\Gamma \geq 1$. It is a different, equally correct, narrative about this sensitivity analysis, but the new narrative involves two sensitivity parameters, $\Lambda \geq 1$ and $\Delta \geq 1$, where $\Gamma = (\Lambda \times \Delta + 1)/(\Lambda + \Delta)$. For instance, a sensitivity analysis at $\Gamma = 1.25$ is the same as a sensitivity analysis at $(\Lambda, \Delta) = (2, 2)$ because $(\Lambda \times \Delta + 1)/(\Lambda + \Delta)$ is $(2 \times 2 + 1)/(2 + 2) = 5/4 = 1.25$. Amplification of a one-dimensional sensitivity analysis generalizes and connects primal, dual, and simultaneous sensitivity analyses [36].

An amplification has two advantages. First, we may understand Γ in a new way, and this is sometimes an aid to interpretation and discussion of a sensitivity analysis. Second, we may perform and report a one-dimensional sensitivity analysis in terms of Γ, yet without reanalyzing the data, we have available a two-dimensional interpretation of this sensitivity analysis. For each value of Γ, there are infinitely many pairs (Λ, Δ) that satisfy the equation $\Gamma = (\Lambda \times \Delta + 1)/(\Lambda + \Delta)$. For instance, $\Gamma = 1.25$ corresponds with $(\Lambda, \Delta) = (2, 2)$, but $\Gamma = 1.25$ also corresponds with $(\Lambda, \Delta) = (1.5, 3.5)$ and $(\Lambda, \Delta) = (3.5, 1.5)$, because $(1.5 \times 3.5 + 1)/(1.5 + 3.5)$ equals 1.25. Indeed, $\Gamma = 1.25$ corresponds with $(\Lambda, \Delta) = (1.25, \infty)$ and $(\Lambda, \Delta) = (\infty, 1.25)$ by taking limits as either Λ or Δ tends to ∞ in the equation $\Gamma = (\Lambda \times \Delta + 1)/(\Lambda + \Delta)$. That is, a numerical value of Γ defines an equivalent curve in terms of (Λ, Δ), but the sensitivity analysis for all (Λ, Δ) on this curve is the same: the same upper bound on the P-value, the same range of point estimates, the same bounds on the confidence interval.

What is the meaning of (Λ, Δ)?

3.6.2 Sensitivity Analysis in Terms of (Λ, Δ)

In Fisher's theory of randomization inference, the inference uses the distribution of treatment assignments, Z_{ij}, created by the experimenter in randomly assigning treatments. In this theory, potential outcomes and covariates, $\mathscr{F}$, are viewed as "fixed." What does "fixed" mean here? Formally, this theory conditions on the values of potential outcomes and observed and unobserved covariates, $\mathscr{F}$. The investigator flipped a fair coin to determine treatment assignment in pair i, $Z_{i1} = 1$ or $Z_{i1} = 0$, with $Z_{i2} = 1 - Z_{i1}$. Because the coin is a fair coin, it pays no attention to the potential outcomes and covariates, $\mathscr{F}$, and given their values, it still comes up heads half the time, tails half the time. The fairness of this coin does not refer to its marginal probability, 1/2, of a heads; rather, fairness refers to what the coin ignores, how conditioning on $\mathscr{F}$ does not affect the coin. If we view the potential outcomes and covariates, $\mathscr{F}$, as fixed, then the fair coin pays no attention to their values and comes up heads half the time. If, instead, we view the potential outcomes and covariates, $\mathscr{F}$, as sampled from a population, then no matter what their distribution in that population, the fair coin pays no attention to their values and comes up heads half the time; so, the conditional distribution of the coin flip, Z_{ij}, given potential outcomes and covariates, $\mathscr{F}$, is still a fair coin flip. In either case—fixed $\mathscr{F}$ or stochastic $\mathscr{F}$— the validity of Fisher's randomization inferences depend upon flipping fair coins to assign treatments—coins that ignore $\mathscr{F}$—not on the origin of these potential outcomes and observed and unobserved covariates. All this was discussed in detail in Chap. 2. In a randomized experiment, $\pi_k = \pi_\ell$, so (3.14) equals $1/2$, and $\Gamma = 1$ in (3.13), so all three terms in (3.15) equal $1/2$.

The sensitivity analysis in terms of Γ measures departures from randomized treatment assignment. It maintains the structure of a randomized experiment, but it allows violations of random assignment, so that π_k may depend upon potential outcomes and observed and unobserved covariates. How large must the deviation from randomized assignment be to alter the qualitative conclusions that the data would support had the data come from a paired randomized experiment? That is the narrative in terms of Γ.

The narrative in terms of (Λ, Δ) is different, yet it produces the same calculations. The narrative in terms of (Λ, Δ) is an alternative interpretation of the same calculations.

The narrative [110, §3] in terms of (Λ, Δ) does not condition on the potential outcomes, (r_{Tij}, r_{Cij}); rather, it expresses the bias in terms of failure to adjust for u_{ij} having adjusted for $\mathbf{x}_{ij}$. It speaks about dependence between two random variables, r_{Cij} and Z_{ij}, as created by matching exactly for $\mathbf{x}_{ij}$ but failing to match also for u_{ij}.

The parameter Λ restricts the probability that the first individual in pair i will be treated, $1 = Z_{i1} = 1 - Z_{i2}$, given observed and unobserved covariates, $(\mathbf{x}_{ij}, u_{ij})$, but not given the potential outcomes, (r_{Tij}, r_{Cij}). It says this probability is at least $1/(1 + \Lambda)$ and at most $\Lambda/(1 + \Lambda)$.

Under a certain semiparametric model defined in the next subsection, the parameter Δ restricts the probability that $r_{Ci1} > r_{Ci2}$ in pair i given observed and

unobserved covariates, $(\mathbf{x}_{ij}, u_{ij})$. It says that this probability is at least $1/(1+\Delta)$ and at most $\Delta/(1+\Delta)$.

A little algebra [110, Proposition 1] then produces the same sensitivity analysis for $(\Lambda, \Delta) \geq (1, 1)$ as for $\Gamma \geq 1$ whenever $\Gamma = (\Lambda \times \Delta + 1)/(\Lambda + \Delta)$.

For the welding data in Sect. 3.5, the upper bound on the P-value testing the null hypothesis H_0 of no effect in Table 3.2 was 0.047 at $\Gamma = 4$. So, a bias of $\Gamma = 4$ is too small to lead to acceptance of H_0 in a 0.05-level test. Because $4 = (7 \times 9 + 1)/(7 + 9)$, a bias of $(\Lambda, \Delta) = (7, 9)$ yields the same upper bound on the P-value and is too small to lead to acceptance of H_0 in a 0.05-level test. In the absence of a treatment effect, a bias of $(\Lambda, \Delta) = (7, 9)$ means that (u_{i1}, u_{i2}) could increase the odds of treatment, $Z_{i1} = 1$, by a factor of $\Lambda = 7$ and the odds of a positive pair difference in outcomes, $Y_i > 0$, by a factor of $\Delta = 9$.

A bias of $\Gamma = 4$ is also equivalent to $(\Lambda, \Delta) = (9, 7)$, $(\Lambda, \Delta) = (5, 19)$, and $(\Lambda, \Delta) = (19, 5)$. Indeed, $\Gamma = 4$ is equivalent to the entire curve of solutions to $\Gamma = (\Lambda \times \Delta + 1)/(\Lambda + \Delta)$.

The same considerations apply to confidence intervals and point estimates. For more about the interpretation of Γ in terms of (Λ, Δ). see [100, Table 9.1].

3.6.3 Precise Meaning of (Λ, Δ)

The current subsection, which may be skipped, defines the two parameters (Λ, Δ) more precisely. We imagine having sampled $(r_{Tij}, r_{Cij}, Z_{ij}, \mathbf{x}_{ij}, u_{ij})$ independently from an infinite population, then paired treated individuals, $Z_{ij} = 1$, to a control, $Z_{ij} = 0$, to form disjoint pairs matched exactly for $\mathbf{x}_{ij}$ alone, so $\mathbf{x}_{i1} = \mathbf{x}_{i2}$ for each pair i. As is often true of models that entail independent samples from an infinite population, this narrative glosses over certain mildly fictional aspects: there are no infinite populations, conditioning on a continuous real-valued covariate (such as age) in a finite population does not pick out a subpopulation but rather a unique individual and so on. It is commonplace and often mathematically tidy to pretend that one can draw a sample of arbitrary size I independently from an infinite population, that conditioning on continuous covariates defines an infinite subpopulation of the infinite population, and so on. An individual who is greatly upset by these routine fictions and glosses can always return to randomization inference and bias defined by Γ, because these notions are rigorously defined in both finite and infinite populations. In any event, the narrative in terms of (Λ, Δ) does make use of a finite independent sample from an infinite population, that is, an independent and identically distributed sample (iid).

The narrative in terms of (Λ, Δ) further assumes that potential outcomes (r_{Tij}, r_{Cij}) and treatments Z_{ij} are conditionally independent given covariates, $(\mathbf{x}_{ij}, u_{ij})$, so that the bias in treatment assignment in pairs matched for $\mathbf{x}_{ij}$ is entirely due to the failure to also match for u_{ij}. Write $V_i = Z_{i1} - Z_{i2} = \pm 1$ and $W_i = r_{Ci1} - r_{Ci2}$; then, any dependence between V_i and W_i is due to the possibility that $u_{i1} \neq u_{i2}$. Write $\mathscr{C} = \{(\mathbf{x}_{ij}, u_{ij}), i = 1, \ldots, I,\ j = 1, 2\}$ for the covariates.

The quantity $V_i \cdot W_i$ is the treated-minus-control difference in potential responses to control, $V_i \cdot W_i = (Z_{i1} - Z_{i2}) \cdot (r_{Ci1} - r_{Ci2})$, and it would equal the treated-minus-control difference, Y_i, in observed responses, $Y_i = (Z_{i1} - Z_{i2}) \cdot (R_{i1} - R_{i2})$, if the null hypothesis of no treatment effect were true, $H_0 : r_{Tij} = r_{Cij}$ for all i, j. If fair coins had determined treatment assignments within pairs, then $\Pr(V_i = 1|\mathscr{C}) = 1/2$ and $\Pr(V_i \cdot W_i|\mathscr{C})$ would be a distribution symmetric about zero; moreover, if H_0 were also true, then $\Pr(Y_i|\mathscr{C})$ would be symmetric about zero. On the other hand, if both $\Pr(V_i|\mathscr{C})$ and $\Pr(W_i|\mathscr{C})$ are influenced by the unmatched (u_{i1}, u_{i2}), then $\Pr(V_i \cdot W_i|\mathscr{C})$ may not be symmetric about zero; so, $\Pr(Y_i|\mathscr{C})$ may not be symmetric about zero even when H_0 is true. The parameter $\Lambda \geq 1$ limits the strength of the association between V_i and (u_{i1}, u_{i2}), while the parameter $\Delta \geq 1$ limits the strength of the association between W_i and (u_{i1}, u_{i2}).

Specifically, in the ith pair, $V_i = Z_{i1} - Z_{i2} = 2Z_{i1} - 1 = \pm 1$ has $1/(1 + \Lambda) \leq \Pr(V_i = 1|\mathscr{C}) \leq \Lambda/(1 + \Lambda)$. So, at first, the parameter Λ appears to be similar to Γ; however, Λ constrains a probability conditioned on $\mathscr{C}$, whereas Γ constrains a probability conditioned on $\mathscr{F}$. Here, $\mathscr{C}$ contains the observed and unobserved covariates, $(\mathbf{x}_{ij}, u_{ij})$, but $\mathscr{F}$ also contains the potential outcomes $(r_{Tij}, r_{Cij}, \mathbf{x}_{ij}, u_{ij})$.

Because W_i is not generally ± 1, the parameter $\Delta \geq 1$ has a slightly different definition. A model is needed that permits $\Pr(W_i|\mathscr{C})$ to exhibit a degree of asymmetry about zero produced by (u_{i1}, u_{i2}) that is controlled by a parameter Δ. Douglas Wolfe [136] proposed a semiparametric family of asymmetric distributions created by taking an arbitrary symmetric distribution and deforming it. In Wolfe's family, the distribution $\Pr(W_i|\mathscr{C})$ satisfies

$$\Pr(W_i \geq w|\mathscr{C}) = \omega_i \cdot \Pr(W_i \leq -w|\mathscr{C}) \text{ for all } w > 0. \tag{3.28}$$

Every distribution symmetric about zero satisfies (3.28) with $\omega_i = 1$, so (3.28) defines a one-parameter family of deformations of symmetry. Expressed differently, Wolfe's family (3.28) places no constraint on the nature of the distribution of $|W_i|$, but the value $W_i = |W_i|$ is ω_i times more likely than the value $W_i = -|W_i|$. As used here, the value of ω_i may vary from one pair i to another i' to reflect differences in (u_{i1}, u_{i2}); see [110, Equation (9)]. Assuming (3.28), the parameter Δ is defined to restrict ω_i by $1/\Delta \leq \omega_i \leq \Delta$ for all i. So, $\Delta = 1$ entails symmetry of W_i for all i. This restriction $1/\Delta \leq \omega_i \leq \Delta$ together with (3.28) implies

$$\frac{1}{1+\Delta} \leq \Pr(W_i > 0 \,|\, \mathscr{C}, \, |W_i| = w) \leq \frac{\Delta}{1+\Delta} \text{ for all } w > 0. \tag{3.29}$$

The claim was that a sensitivity analysis defined by (Λ, Δ) yields the same sensitivity bounds as $\Gamma = (\Lambda \times \Delta + 1)/(\Lambda + \Delta)$. The proof of this [110, Proposition 1] essentially shows that the probability that pair i will make a positive contribution to, say, Wilcoxon's signed rank statistic T is sharply bounded above by $(\Lambda \times \Delta + 1)/(\Lambda + \Delta)$, or more precisely that $\Pr(V_i \cdot W_i > 0 \,|\, \mathscr{C}, \, |W_i|) \leq (\Lambda \times \Delta + 1)/(\Lambda + \Delta)$. In other words, two sensitivity analyses—one defined in

terms of Γ, the other in terms of (Λ, Δ)—produce identical sensitivity bounds. Hence, that one sensitivity analysis has two equivalent interpretations.

The assumptions underlying the amplification may be weakened, at the risk of making these assumptions somewhat more obscure. Wolfe's [136] model (3.28) implies $\Pr(W_i > 0 \,|\, \mathscr{C}, |W_i| = w) = \omega_i/(1 + \omega_i)$ for all $w > 0$, but equality is not needed here; rather, the less restrictive inequality (3.29) is needed. The function $\mathrm{abz}(w) = \Pr(W > 0 \,|\, \mathscr{C}, |W| = w)$ in (3.29) was introduced by Albers, Bickel, and van Zwet [2] in their study of the distributions of signed rank statistics, and it will reappear in Chap. 19. For various distributions, $\mathrm{abz}(w)$ is plotted against w in Figure 3 of [98, §3]. Under Wolfe's model (3.28), $\mathrm{abz}(w)$ is a constant not varying with w, but the argument leading to the amplification requires only that $\mathrm{abz}(w)$ be bounded in (3.29). So, instead of assuming Wolfe's model (3.28), one could assume the much weaker conclusion (3.29) that follows from Wolfe's model. For example, if a positive constant is added to a double-exponential or logistic random variable, the resulting distribution is not symmetric about zero, but it also does not satisfy Wolfe's model (3.28); nonetheless, it does satisfy (3.29) for some $\Delta < \infty$, essentially because, for these distributions, $\mathrm{abz}(w)$ is bounded strictly below one.[17]

The parameter Γ is not restricted to matched pairs: it is applicable in matched samples with multiple controls [97] or in unmatched stratified comparisons [101, 111]. In all cases, we may interpret a value of Γ in terms of (Λ, Δ) by describing the impact of a bias of magnitude Γ on a matched pair.

3.7 Bias Due to Incomplete Matching

Chapter 3 has followed Chap. 2 in focusing on internal validity, as discussed in Sect. 2.6. That is, the focus has been on inference about treatment effects for the $2I$ matched individuals, not on whether the same treatment effects would be found in the population of L individuals. When treatment effects vary from person to person, as seemed to be the case in Sect. 2.5, changing the individuals under study may change the magnitude of the effect.

Although not required for internal validity, it is common in practice to match all of the treated subjects in the population, so that the number of matched pairs, I, equals the number of treated subjects $\sum Z_\ell$ in the available population of L individuals. The goal here is an aspect of external validity, specifically the ability to speak about treatment effects in the original population of L individuals. If the population of L individuals were itself a random sample from an infinite population, and if all treated subjects were matched, then under the naïve model (3.5)–(3.8), the average treated-minus-control difference in observed responses, $(1/I)\sum Y_i$,

[17] In contrast, for a Normal distribution with positive expectation, $\mathrm{abz}(w) \to 1$ as $w \to \infty$, so (3.29) does not hold for any $\Delta < \infty$. Looking at Figure 3 of [98, §3], this fact seems to be a peculiarity of the Normal distribution with its exceptionally thin tails, not a peculiarity of (3.29).

would be unbiased for the expected treatment effect on people who typically receive the treatment, namely $E(r_T - r_C \mid Z = 1)$; however, this is typically untrue if some treated subjects are deleted, creating what is known as the "bias due to incomplete matching" [108]. Unless explicitly instructed to do otherwise, the matching methods in Part II will match all treated subjects.

If some treated subjects, ℓ, in the population have propensity scores near 1, $e(\mathbf{x}_\ell) \approx 1$, they will be very difficult to match. Virtually everyone with this $\mathbf{x}_\ell$ will receive treatment. Rather than delete individuals one at a time based on extreme propensity scores, $e(\mathbf{x}_\ell)$, it is usually better to go back to the covariates themselves, $\mathbf{x}_\ell$, perhaps redefining the population under study to be a subpopulation of the original population of L subjects. A population defined in terms of $e(\mathbf{x}_\ell)$ is likely to have little meaning to other investigators, whereas a population defined in terms of one or two familiar covariates from $\mathbf{x}_\ell$ will have a clear meaning. In this redefined population, all treated subjects are matched. For instance, in a study of the effects of joining a gang at age 14 on subsequent violence [44, 45], a handful of the most extremely violent, chronically violent boys at age 13 all joined gangs at age 14. These few extremely violent boys had no relevant controls—all of the potential controls were substantially less violent than these few boys prior to age 14. Disappointing as this may be, there is no plausible way to estimate the effect of a treatment in a subpopulation that always receives the treatment. Using violence prior to age 14, the study population was redefined to exclude the extremely chronically violent subpopulation, with no claim that similar effects would be found in that subpopulation. See Sect. 13.4 for further discussion of the gang study.

In medical research, it is common for a clinical trial to begin with various inclusion and exclusion criteria defined in terms of a few pretreatment covariates. In parallel, Colin Fogarty et al. [30] proposed a computerized method that helps construct an interpretable subpopulation of treated individuals when it is not possible to match all treated individuals. The method seeks to define the subpopulation in terms of a few covariates from $\mathbf{x}_\ell$ rather than from the propensity score, $e(\mathbf{x}_\ell)$. See also [21, 99, 127] for various methods of subset matching.

3.8 Summary

Two simple models for treatment assignment in observational studies have been discussed. The two models define and divide the two tasks faced by the investigator.

The first model is naïve: it says that two people who look comparable in terms of observed covariates $\mathbf{x}_\ell$ are comparable. People who look comparable in terms of $\mathbf{x}_\ell$ are said to be "ostensibly comparable"; on the surface, they appear to be comparable, but they may not be so. If the naïve model were true, it would suffice to match treated and control subjects for the observed covariates $\mathbf{x}_\ell$. More precisely, if the naïve model were true, and treated and control subjects were matched so that $\mathbf{x}_{ij} = \mathbf{x}_{ik}$ for different subjects j and k in the same matched set i, then this alone would reproduce the distribution of treatment assignments in a randomized experiment.

The key difficulty in an observational study is that there is usually little or no reason to believe the naïve model is true.

The sensitivity analysis model says people who look comparable in terms of observed covariates $\mathbf{x}_\ell$ may differ in terms of one or more covariates u_ℓ that were not measured. The sensitivity analysis model says that two subjects, j and k, who look comparable in terms of observed covariates, and so might be placed in the same matched set i—that is, two subjects with $\mathbf{x}_{ij} = \mathbf{x}_{ik}$—may differ in their odds of treatment by a factor of $\Gamma \geq 1$. When $\Gamma = 1$, the sensitivity analysis model reduces to the naïve model, yielding the distribution of treatment assignments in a randomized experiment. When $\Gamma > 1$, the treatment assignment probabilities are unknown, but unknown to a bounded degree. For each fixed value of $\Gamma \geq 1$, there is a range of possible inferences, for instance, an interval of possible P-values or point estimates or endpoints for a confidence interval. For $\Gamma = 1$, this interval of P-values or estimates reduces to a single point, namely the randomization inference. As $\Gamma \to \infty$, the interval widens until, at some point, it is so long as to be uninformative, for instance, including both small and large P-values. The sensitivity analysis determines the magnitude of bias, measured by Γ, that would need to be present to qualitatively alter the conclusions of the study, that is, to produce an interval so long that it is uninformative. There is always such a Γ, but the numerical value of Γ varies dramatically from one observational study to the next. The naïve model assumes $\Gamma = 1$. The phrase "association does not imply causation" refers to letting $\Gamma \to \infty$. The sensitivity analysis determines the value of Γ that is relevant in light of the data at hand.

The two models define and divide the two tasks in an observational study. The first task is to compare people who look comparable. The first task can be done somewhat mechanically and completed: we may reach a stage where we are all forced to agree that the people who are being compared under alternative treatments do indeed look comparable in terms of observed covariates $\mathbf{x}$. At this point, the first task is done. Part II of this book discusses the first task, namely matching for observed covariates. The second task is to address the possibility that differing outcomes in treated and control groups are not effects caused by the treatment, but instead reflect some way in which treated and control groups are not comparable in terms of covariates that were not measured. The second task is not a mechanical task, not one that could be handed over to a computer. The second task, being attended by controversy, is more challenging, and hence more interesting. Sensitivity analysis, the tactics in Chap. 5, and the concepts in Parts III–V are aimed at the second task.

3.9 Further Reading

In this book, see Chap. 9 for further discussion of propensity scores, and see Chap. 5 and Parts III–V for further discussion of sensitivity analysis. Propensity scores originate in [104] and are discussed in [22, 24, 41, 46, 48, 52, 63, 65, 67, 78, 79,

81, 85, 94, 106, 107, 113–115, 130], [93, Chapter 3], and [100, Chapter 5]. The method of sensitivity analysis described here is not restricted to matched pairs and is discussed in [28, 29, 31, 37, 83, 88, 89, 94, 97, 101, 103, 140], [93, Chapter 4], and [100, Chapter 9]; see also [59]. Amplification is discussed in [110] and [100, Chapter 9]. Qingyuan Zhao defines the value of Γ that yields P-value α, conventionally $\alpha = 0.05$, as the "sensitivity value"; it is a random variable, and he studies its stochastic properties. Other approaches to sensitivity analysis are discussed in [16, 18–20, 35, 36, 38, 53, 61, 64–66, 70, 80, 82, 105, 105, 116, 121, 138, 142]. For a few applications of sensitivity analysis, see [6, 17, 25, 60, 71, 122, 143].

3.10 Software

The `senWilcox` function in the `DOS2` package in `R` performs the sensitivity analysis for Wilcoxon's signed rank test, the Hodges–Lehmann estimate, and the associated confidence interval. Also, there is a `stata` module `RBOUNDS` by Markus Gangl and an `R` package `rbounds` by Luke Keele. See also the software section in Chap. 19 for sensitivity analyses for other test statistics.

For means and other m-statistics, sensitivity analysis is implemented in the `senm` and `senmCI` functions of the `sensitivitymult` package. The shinyapp mentioned in the Preface runs `sensitivitymult` in the background. These functions may also be used when matching to multiple controls, perhaps a variable number of controls. The package `sensitivityfull` performs sensitivity analysis for full matching.

For unmatched comparisons, with or without strata, the `R` package `senstrat` performs a sensitivity analysis with various test statistics [101].

The `amplify` function in the `sensitivitymult` and `sensitivitymv` packages in `R` performs the simple calculations required to amplify one sensitivity parameter Γ into two equivalent parameters Λ and Δ as in [110].

Samuel Pimentel's [75, §4.3] `R` package `rcbsubset` implements a version of subset matching in Sect. 3.7.

The exact sensitivity analysis in the Appendix is implemented in the `senWilcoxExact` function in the `R` package `DOS2`.

3.11 Data

The welder data from [134] in Sect. 3.5 is in the `R` package `DOS2` as `werfel` whose help-file reproduces aspects of Sect. 3.5.

Appendix: Exact Computations for Sensitivity Analysis

For small to moderate I, the exact upper bound on the distribution of Wilcoxon's signed rank statistic may be obtained quickly in R as the convolution of I probability generating functions. Only the situation without ties is considered. Pagano and Tritchler [73] observed that permutation distributions may often be obtained in polynomial time by applying the fast Fourier transform to the convolution of characteristic functions or generating functions. Here, probability generating functions are used along with the R function `convolve`.

The distribution of $\overline{\overline{T}}$ is the distribution of the sum of I independent random variables, $i = 1, 2, \ldots, I$, taking the value i with probability $\Gamma/(1+\Gamma)$ and the value 0 with probability $1/(1+\Gamma)$. The ith random variable has probability generating function

$$h_i(x) = \frac{1}{1+\Gamma} + \frac{\Gamma x^i}{1+\Gamma}, \tag{3.30}$$

and $\overline{\overline{T}}$ has generating function $\Pi_{i=1}^{I} h_i(x)$. In R, the generating function of a random variable taking integer values 0, 1, ..., B is represented by a vector of dimension $B+1$ whose $b+1$ coordinate gives the probability that the random variable equals b. For instance, $h_3(x)$ is represented by

$$\left(\frac{1}{1+\Gamma},\ 0,\ 0,\ \frac{\Gamma}{1+\Gamma}\right). \tag{3.31}$$

The distribution of $\overline{\overline{T}}$ is obtained by convolution as a vector with $1 + I(I+1)/2$ coordinates representing $\Pi_{i=1}^{I} h_i(x)$ and giving $\Pr\left(\overline{\overline{T}} = b\right)$ for $b = 0, 1, \ldots, I(I+1)/2$, where $I(I+1)/2 = \sum i$.

Bibliography

1. Abadie, A.: Semiparametric difference-in-differences estimators. Rev. Econ. Stud. **72**, 1–19 (2005)
2. Albers, W., Bickel, P.J., van Zwet, W.R.:Asymptotic expansions for the power of distribution free tests in the one-sample problem. Ann. Stat. **4**, 108–156 (1976)
3. Angrist, J., Hahn, J.: When to control for covariates? Panel asymptotics for estimates of treatment effects. Rev. Econ. Stat. **86**, 58–72 (2004)
4. Angrist, J.D., Krueger, A.B.: Empirical strategies in labor economics. In: Ashenfelter, O., Card, D. (eds.) Handbook of Labor Economics, vol. 3, pp. 1277–1366. Elsevier, New York (1999)
5. Angrist, J.D., Lavy, V.: Using Maimonides' rule to estimate the effect of class size on scholastic achievement. Q. J. Econ. **114**, 533–575 (1999)

6. Armstrong, C.S., Blouin, J.L., Larcker, D.F.:The incentives for tax planning. J. Accounting Econ. **53**, 391–411 (2012)
7. Athey, S., Imbens, G.W.:Identification and inference in nonlinear difference-in-differences models. Econometrica **74**, 431–497 (2006)
8. Becker, S.O., Caliendo, M.:Sensitivity analysis for average treatment effects. Stata J. **7**, 71–83 (2007)
9. Bertrand, M., Duflo, E., Mullainathan, S.:How much should we trust difference-in-differences estimates? Q. J. Econ. **119**, 249–275 (2004)
10. Besley, T., Case, A.: Unnatural experiments? Estimating the incidence of endogenous policies. Econ. J. **110**, 672–694 (2000)
11. Bross, I.D.J.: Statistical criticism. Cancer **13**, 394–400 (1961)
12. Bross, I.D.J.: Spurious effects from an extraneous variable. J. Chron. Dis. **19**, 637–647 (1966)
13. Campbell, D.T.: Factors relevant to the validity of experiments in social settings. Psychol. Bull. **54**, 297–312 (1957)
14. Campbell, D.T.: Reforms as experiments. Am. Psychol. **24**, 409–429 (1969)
15. Campbell, D.T.: Methodology and Epistemology for Social Science: Selected Papers. University of Chicago Press, Chicago (1988)
16. Carnegie, N.B., Harada, M. Hill, J.L.: Assessing sensitivity to unmeasured confounding using a simulated potential confounder. J. Res. Educ. Effect **9**, 395–420 (2016)
17. Chi, S.S., Shanthikumar, D.M.: Local bias in Google search and the market response around earnings announcements. Account Rev. **92**, 115–143 (2016)
18. Copas, J.B., Eguchi, S.: Local sensitivity approximations for selectivity bias. J. R. Stat. Soc. B **63**, 871–896 (2001)
19. Copas, J.B., Li, H.G.: Inference for non-random samples. J. R. Stat. Soc. B **59**, 55–77 (1997)
20. Cornfield, J., Haenszel, W., Hammond, E., Lilienfeld, A., Shimkin, M., Wynder, E.: Smoking and lung cancer: recent evidence and a discussion of some questions. J. Natl. Cancer Inst. **22**, 173–203 (1959)
21. Crump, R.K., Hotz, V.J., Imbens, G.W., Mitnik, O.A.:Dealing with limited overlap in estimation of average treatment effects. Biometrika **96**, 187–199 (2009)
22. D'Agostino, R.B.: Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. Stat. Med. **17**, 2265–2281 (1998)
23. Dawid, A.P.: Conditional independence in statistical theory (with Discussion). J. R. Stat. Soc. B **41**, 1–31 (1979)
24. Dehejia, R.H., Wahba, S.: Propensity score-matching methods for nonexperimental causal studies. Rev. Econ. Stat. **84**, 151–161 (2002)
25. Diprete, T.A., Gangl, M.: Assessing bias in the estimation of causal effects: Rosenbaum bounds on matching estimators and instrumental variables estimating with imperfect instruments. Sociol. Method **34**, 271–310 (2004)
26. Dynarski, S.M.: Does aid matter? Measuring the effect of student aid on college attendance and completion. Am. Econ. Rev. **93**, 279–288 (2003)
27. Fenech, M., Changb, W.P., Kirsch-Voldersc, M., Holland, N., Bonassie, S., Zeiger, E.: HUMN project: detailed description of the scoring criteria for the cytokinesis-block micronucleus assay using isolated human lymphocyte cultures. Mutat. Res. **534**, 65–75 (2003)
28. Fogarty, C.B.: Studentized sensitivity analysis for the sample average treatment effect in paired observational studies. J. Am. Stat. Assoc. (2019, to appear). https://doi.org/10.1080/01621459.2019.1632072
29. Fogarty, C.B., Small, D.S.: Sensitivity analysis for multiple comparisons in matched observational studies through quadratically constrained linear programming. J. Am. Stat. Assoc. **111**, 1820–1830 (2016)
30. Fogarty, C.B., Mikkelsen, M.E., Gaieski, D.F., Small, D.S.:Discrete optimization for interpretable study populations and randomization inference in an observational study of severe sepsis mortality. J. Am. Stat. Assoc. **111**, 447–458 (2016)

31. Fogarty, C.B., Shi, P., Mikkelsen, M.E., Small, D.S.:Randomization inference and sensitivity analysis for composite null hypotheses with binary outcomes in matched observational studies. J. Am. Stat. Assoc. **112**, 321–331 (2017)
32. Foster, E.M., Bickman, L.: Old wine in new skins: the sensitivity of established findings to new methods. Eval. Rev. **33**, 281–306 (2009)
33. Frangakis, C.E., Rubin, D.B.: Principal stratification in causal inference. Biometrics **58**, 21–29 (2002)
34. Franklin, J.M., Eddings, W., Glynn, R.J., Schneeweiss, S.: Regularized regression versus the high-dimensional propensity score for confounding adjustment in secondary database analyses. Am. J. Epidemiol. **182**, 651–657 (2015)
35. Gastwirth, J.L.: Methods for assessing the sensitivity of comparisons in Title VII cases to omitted variables. Jurimetrics J. **33**, 19–34 (1992)
36. Gastwirth, J.L., Krieger, A.M., Rosenbaum, P.R.: Dual and simultaneous sensitivity analysis for matched pairs. Biometrika **85**, 907–920 (1998)
37. Gastwirth, J.L., Krieger, A.M., Rosenbaum, P.R.: Asymptotic separability in sensitivity analysis. J. R. Stat. Soc. B **62**, 545–555 (2000)
38. Greenland, S.: Basic methods of sensitivity analysis. Int. J. Epidemiol. **25**, 1107–1116 (1996)
39. Greevy, R., Lu, B., Silber, J.H., Rosenbaum, P.R.: Optimal matching before randomization. Biostatistics **5**, 263–275 (2004)
40. Gross, D.B., Souleles, N.S.: Do liquidity constraints and interest rates matter for consumer behavior? Evidence from credit card data. Q. J. Econ. **117**, 149–185 (2002)
41. Hahn, J.Y.: On the role of the propensity score in efficient semiparametric estimation of average treatment effects. Econometrica **66**, 315–331 (1998)
42. Hamermesh, D.S.: The craft of labormetrics. Ind. Labor Relat. Rev. **53**, 363–380 (2000)
43. Hansen, B.B.: The prognostic analogue of the propensity score. Biometrika **95**, 481–488 (2008)
44. Haviland, A., Nagin, D.S., Rosenbaum, P.R.: Combining propensity score matching and group-based trajectory analysis in an observational study. Psychol. Methods **12**, 247–267 (2007)
45. Haviland, A.M., Nagin, D.S., Rosenbaum, P.R., Tremblay, R.E.: Combining group-based trajectory modeling and propensity score matching for causal inferences in nonexperimental longitudinal data. Dev. Psychol. **44**, 422–436 (2008)
46. Hirano, K., Imbens, G.W., Ridder, G.: Efficient estimation of average treatment effects using the estimated propensity score. Econometrica **71**, 1161–1189 (2003)
47. Hill, A.B.: The environment and disease: association or causation? Proc. R. Soc. Med. **58**, 295–300 (1965)
48. Hill, J.L., Waldfogel, J., Brooks-Gunn, J., Han, W.J.: Maternal employment and child development: a fresh look using newer methods. Dev. Psychol. **41**, 833–850 (2005)
49. Ho, D.E., Imai, K., King, G., Stuart, E.A.: Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. Polit. Anal. **15**, 199–236 (2007)
50. International Agency for Research on Cancer: IARC Monographs on the Valuation of Carcinogenic Risks of Chemicals to Humans: Chromium, Nickel and Welding, vol. 49, pp. 447–525. IARC, Lyon (1990)
51. Imai, K.: Statistical analysis of randomized experiments with non-ignorable missing binary outcomes: an application to a voting experiment. Appl. Stat. **58**, 83–104 (2009)
52. Imbens, G.W.: The role of the propensity score in estimating dose response functions. Biometrika **87**, 706–710 (2000)
53. Imbens, G.W.: Sensitivity to exogeneity assumptions in program evaluation. Am. Econ. Rev. **93**, 126–132 (2003)
54. Imbens, G.W.: Nonparametric estimation of average treatment effects under exogeneity: a review. Rev. Econ. Stat. **86**, 4–29 (2004)
55. Imbens, G.W., Wooldridge, J.M.: Recent developments in the econometrics of program evaluation. J. Econ. Lit. **47**, 5–86 (2009)

56. Joffe, M.M., Ten Have, T.R., Feldman, H.I., Kimmel, S.E.: Model selection, confounder control, and marginal structural models: review and new applications. Am Stat. **58**, 272–279 (2004)
57. Johnson, B.A., Tsiatis, A.A.: Estimating mean response as a function of treatment duration in an observational study, where duration may be informatively censored. Biometrics **60**, 315–323 (2004)
58. Katan, M.B.: Commentary: Mendelian randomization, 18 years on. Int. J. Epidemiol. **33**, 10–11 (2004)
59. Keele, L.J.: Rbounds: an R package for sensitivity analysis with matched data. http://www.polisci.ohio-state.edu/faculty/lkeele/rbounds.html
60. Lee, M.J., Lee, S.J.: Sensitivity analysis of job-training effects on reemployment for Korean women. Empir. Econ. **36**, 81–107 (2009)
61. Lin, D.Y., Psaty, B.M., Kronmal, R.A.: Assessing sensitivity of regression to unmeasured confounders in observational studies. Biometrics **54**, 948–963 (1998)
62. Lu, B., Rosenbaum, P.R.: Optimal matching with two control groups. J. Comput. Graph Stat. **13**, 422–434 (2004)
63. Manski, C.: Nonparametric bounds on treatment effects. Am. Econ. Rev. **80**, 319–323 (1990)
64. Manski, C.F.: Identification Problems in the Social Sciences. Harvard University Press, Cambridge (1995)
65. Manski, C.F., Nagin, D.S.: Bounding disagreements about treatment effects: a case study of sentencing and recidivism. Sociol. Method **28**, 99–137 (1998)
66. Marcus, S.M.: Using omitted variable bias to assess uncertainty in the estimation of an AIDS education treatment effect. J. Educ. Behav. Stat. **22**, 193–201 (1997)
67. McCaffrey, D.F., Ridgeway, G., Morral, A.R.: Propensity score estimation with boosted regression for evaluating causal effects in observational studies. Psychol. Methods **9**, 403–425 (2004)
68. McKillip, J.: Research without control groups: a control construct design. In: Bryant, F.B., et al. (eds.) Methodological Issues in Applied Social Psychology, pp. 159–175. Plenum Press, New York (1992)
69. Meyer, B.D.: Natural and quasi-experiments in economics. J. Bus. Econ. Stat. **13**, 151–161 (1995)
70. Mitra, N. Heitjan, D.F.: Sensitivity of the hazard ratio to nonignorable treatment assignment in an observational study. Stat. Med. **26**, 1398–1414 (2007)
71. Normand, S-L., Landrum, M.B., Guadagnoli, E., Ayanian, J.Z., Ryan, T.J., Cleary, P.D., McNeil, B.J.: Validating recommendations for coronary angiography following acute myocardial infarction in the elderly: a matched analysis using propensity scores. J. Clin. Epidemiol. **54**, 387–398 (2001)
72. Normand, S-L., Sykora, K., Li, P., Mamdani, M., Rochon, P.A., Anderson, G.M.: Readers guide to critical appraisal of cohort studies: 3. Analytical strategies to reduce confounding. Br. Med. J. **330**, 1021–1023 (2005)
73. Pagano, M., Tritchler, D.: On obtaining permutation distributions in polynomial time. J. Am. Stat. Assoc. **78**, 435–440 (1983)
74. Peel, M.J., Makepeace, G.H.: Differential audit quality, propensity score matching and Rosenbaum bounds for confounding variables. J. Bus. Financ. Account **39**, 606–648 (2012)
75. Pimentel, S.D.: Large, sparse optimal matching with R package rcbalance. Obs. Stud. **2**, 4–23 (2016)
76. Reynolds, K.D., West, S.G.: A multiplist strategy for strengthening nonequivalent control group designs. Eval. Rev. **11**, 691–714 (1987)
77. Richardson, A., Hudgens, M.G., Gilbert, P.B., Fine, J.P.: Nonparametric bounds and sensitivity analysis of treatment effects. Stat. Sci. **29**, 596–618 (2014)
78. Robins, J.M., Ritov, Y.: Toward a curse of dimensionality appropriate (CODA) asymptotic theory for semi-parametric models. Stat. Med. **16**, 285–319 (1997)
79. Robins, J.M., Mark, S.D., Newey, W.K.: Estimating exposure effects by modeling the expectation of exposure conditional on confounders. Biometrics **48**, 479–495 (1992)

80. Robins, J.M., Rotnitzky, A., Scharfstein, D.: Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In: Halloran, E., Berry, D. (eds.) Statistical Models in Epidemiology, pp. 1–94. Springer, New York (1999)
81. Rosenbaum, P.R.: Conditional permutation tests and the propensity score in observational studies. J. Am. Stat. Assoc. **79**, 565–574 (1984)
82. Rosenbaum, P.R.: Dropping out of high school in the United States: an observational study. J. Educ. Stat. **11**, 207–224 (1986)
83. Rosenbaum, P.R.: Sensitivity analysis for certain permutation inferences in matched observational studies. Biometrika **74**, 13–26 (1987)
84. Rosenbaum, P.R.: The role of a second control group in an observational study (with Discussion). Stat. Sci. **2**, 292–316 (1987)
85. Rosenbaum, P.R.: Model-based direct adjustment. J. Am. Stat. Assoc. **82**, 387–394 (1987)
86. Rosenbaum, P.R.: The role of known effects in observational studies. Biometrics **45**, 557–569 (1989)
87. Rosenbaum, P.R.: On permutation tests for hidden biases in observational studies. Ann. Stat. **17**, 643–653 (1989)
88. Rosenbaum, P.R.: Hodges-Lehmann point estimates in observational studies. J. Am. Stat. Assoc. **88**, 1250–1253 (1993)
89. Rosenbaum, P.R.: Quantiles in nonrandom samples and observational studies. J. Am. Stat. Assoc. **90**, 1424–1431 (1995)
90. Rosenbaum, P.R.: Signed rank statistics for coherent predictions. Biometrics **53**, 556–566 (1997)
91. Rosenbaum, P.R.: Choice as an alternative to control in observational studies (with Discussion). Stat. Sci. **14**, 259–304 (1999)
92. Rosenbaum, P.R.: Stability in the absence of treatment. J. Am. Stat. Assoc. **96**, 210–219 (2001)
93. Rosenbaum, P.R.: Observational Studies, 2nd edn. Springer, New York (2002)
94. Rosenbaum, P.R.: Covariance adjustment in randomized experiments and observational studies (with Discussion). Stat. Sci. **17**, 286–327 (2002)
95. Rosenbaum, P.R.: Design sensitivity in observational studies. Biometrika **91**, 153–164 (2004)
96. Rosenbaum, P.R.: Heterogeneity and causality: unit heterogeneity and design sensitivity in observational studies. Am Stat. **59**, 147–152 (2005)
97. Rosenbaum, P.R.: Sensitivity analysis for m-estimates, tests, and confidence intervals in matched observational studies. Biometrics **63**, 456–464 (2007)
98. Rosenbaum, P.R.: Design sensitivity and efficiency in observational studies. J. Am. Stat. Assoc. **105**, 692–702 (2010)
99. Rosenbaum, P.R.: Optimal matching of an optimally chosen subset in observational studies. J. Comput. Graph Stat. **21**, 57–71 (2012)
100. Rosenbaum, P.R.: Observation and Experiment: An Introduction to Causal Inference. Harvard University Press, Cambridge, MA (2017)
101. Rosenbaum, P.R.: Sensitivity analysis for stratified comparisons in an observational study of the effect of smoking on homocysteine levels. Ann. Appl. Stat. **12**, 2312–2334 (2018)
102. Rosenbaum, P.R.: Modern algorithms for matching in observational studies. Ann. Rev. Stat. Appl. **7**, 143–176 (2020)
103. Rosenbaum, P.R., Krieger, A.M.: Sensitivity analysis for two-sample permutation inferences in observational studies. J. Am. Stat. Assoc. **85**, 493–498 (1990)
104. Rosenbaum, P.R., Rubin, D.B.: The central role of the propensity score in observational studies for causal effects. Biometrika **70**, 41–55 (1983)
105. Rosenbaum, P.R., Rubin, D.B.: Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. J. R. Stat. Soc. B **45**, 212–218 (1983)
106. Rosenbaum, P.R., Rubin, D.B.: Reducing bias in observational studies using subclassification on the propensity score. J. Am. Stat. Assoc. **79**, 516–524 (1984)
107. Rosenbaum, P.R., Rubin, D.B.: Constructing a control group by multivariate matched sampling methods that incorporate the propensity score. Am. Stat. **39**, 33–38 (1985)

108. Rosenbaum, P.R., Rubin, D.B.: The bias due to incomplete matching. Biometrics **41**, 106–116 (1985)
109. Rosenbaum, P.R., Silber, J.H.: Sensitivity analysis for equivalence and difference in an observational study of neonatal intensive care units. J. Am. Stat. Assoc. **104**, 501–511 (2009)
110. Rosenbaum, P.R., Silber, J.H.: Amplification of sensitivity analysis in observational studies. J. Am. Stat. Assoc. **104**, 1398–1405 (2009)
111. Rosenbaum, P.R., Small, D.S.: An adaptive Mantel-Haenszel test for sensitivity analysis in observational studies. Biometrics **73**, 422–430 (2017)
112. Rosenzweig, M.R., Wolpin, K.I.: Natural 'natural experiments' in economics. J. Econ. Lit. **38**, 827–874 (2000)
113. Rotnitzky, A., Robins, J.M.: Semiparametric regression estimation in the presence of dependent censoring. Biometrika **82**, 805–820 (1995)
114. Rubin, D.B., Thomas, N.: Characterizing the effect of matching using linear propensity score methods with normal distribution. Biometrika **79**, 797–809 (1992)
115. Rubin, D.B., Thomas, N.: Combining propensity score matching with additional adjustments for prognostic covariates. J. Am. Stat. Assoc. **95**, 573–585 (2000)
116. Rudolph, K.E., Stuart, E.A.: Using sensitivity analyses for unobserved confounding to address covariate measurement error in propensity score methods. Am. J. Epidemiol. **187**, 604–613 (2017)
117. Rutter, M.: Identifying the Environmental Causes of Disease: How do We Decide What to Believe and When to Take Action? Academy of Medical Sciences, London (2007)
118. Shadish, W.R., Cook, T.D.: The renaissance of field experimentation in evaluating interventions. Annu. Rev. Psychol. **60**, 607–629 (2009)
119. Shadish, W.R., Cook, T.D., Campbell, D.T.: Experimental and Quasi-Experimental Designs for Generalized Causal Inference. Houghton-Mifflin, Boston (2002)
120. Shepherd, B.E., Gilbert, P.B., Jemiai, Y., Rotnitzky, A.: Sensitivity analyses comparing outcomes only existing in a subset selected post-randomization, conditional on covariates, with application to HIV vaccine trials. Biometrics **62**, 332–342 (2006)
121. Shepherd, B.E., Gilbert, P.B., Mehrotra, D.V.: Eliciting a counterfactual sensitivity parameter. Am Stat. **61**, 56–63 (2007)
122. Silber, J.H., Rosenbaum, P. R., Trudeau, M.E., Chen, W., Zhang, X., Lorch, S.L., Rapaport-Kelz, R., Mosher, R.E., Even-Shoshan, O.: Preoperative antibiotics and mortality in the elderly. Ann. Surg. **242**, 107–114 (2005)
123. Small, D., Rosenbaum, P.R.: War and wages: The strength of instrumental variables and their sensitivity to unobserved biases. J. Am. Stat. Assoc. **103**, 924–933 (2008)
124. Spielman, R.S., Ewens, W.J.: A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test. Am. J. Hum. Genet. **62**, 450–458 (1998)
125. Stolley, P.D.: When genius errs—R.A. Fisher and the lung cancer controversy. Am. J. Epidemiol. **133**, 416–425 (1991)
126. Stone, R.: The assumptions on which causal inferences rest. J. R. Stat. Soc. B **55**, 455–466 (1993)
127. Traskin, M., Small, D.S.: Defining the study population for an observational study to ensure sufficient overlap: a tree approach. Stat. Biosci. **3**, 94–118 (2011)
128. Trochim, W.M.K.: Pattern matching, validity and conceptualization in program evaluation. Eval. Rev. **9**, 575–604 (1985)
129. Vandenbroucke, J.P.: When are observational studies as credible as randomized trials? Lancet **363**, 1728–1731 (2004)
130. VanderWeele, T.: The use of propensity score methods in psychiatric research. Int. J. Methods Psychol. Res. **15**, 95–103 (2006)
131. Volpp, K.G., Rosen, A.K., Rosenbaum, P.R., Romano, P.S., Even-Shoshan, O., Wang, Y., Bellini, L., Behringer, T., Silber, J.H.: Mortality among hospitalized Medicare beneficiaries in the first 2 years following ACGME resident duty hour reform. J. Am. Med. Assoc. **298**, 975–983 (2007)

132. Wang, L.S., Krieger, A.M.: Causal conclusions are most sensitive to unobserved binary covariates. Stat. Med. **25**, 2257–2271 (2006)
133. Weiss, N.S.: Can the "specificity" of an association be rehabilitated as a basis for supporting a causal hypothesis? Epidemiology **13**, 6–8 (2002)
134. Werfel, U., Langen, V., Eickhoff, I., Schoonbrood, J., Vahrenholz, C., Brauksiepe, A., Popp, W., Norpoth, K.: Elevated DNA single-strand breakage frequencies in lymphocytes of welders. Carcinogenesis **19**, 413–418 (1998)
135. Wintemute, G.J., Wright, M.A., Drake, C.M., Beaumont, J.J.: Subsequent criminal activity among violent misdemeanants who seek to purchase handguns: risk factors and effectiveness of denying handgun purchase. J. Am. Med. Assoc. **285**, 1019–1026 (2001)
136. Wolfe, D.A.: A characterization of population weighted symmetry and related results. J. Am. Stat. Assoc. **69**, 819–822 (1974)
137. Wyss, R., Schneeweiss, S., van der Laan, M., Lendle, S.D., Ju, C., Franklin, J.M.: Using super learner prediction modeling to improve high-dimensional propensity score estimation. Epidemiology **29**, 96–106 (2018)
138. Yu, B.B., Gastwirth, J.L.: Sensitivity analysis for trend tests: application to the risk of radiation exposure. Biostatistics **6**, 201–209 (2005)
139. Zanutto, E., Lu, B., Hornik, R.: Using propensity score subclassification for multiple treatment doses to evaluate a national antidrug media campaign. J. Educ. Behav. Stat. **30**, 59–73 (2005)
140. Zhao, Q.: On sensitivity value of pair-matched observational studies. J. Am. Stat. Assoc. **114**, 713–722 (2019)
141. Zhao, Q.: Covariate balancing propensity score by tailored loss functions. Ann. Stat. **47**, 965–993 (2019)
142. Zhao, Q., Small, D.S., Bhattacharya, B.B.: Sensitivity analysis for inverse probability weighting estimators via the percentile bootstrap. J. R. Stat. Soc. B **81**, 735–761 (2019)
143. Zubizarreta, J.R., Cerda, M., Rosenbaum, P.R.: Effect of the 2010 Chilean earthquake on posttraumatic stress: reducing sensitivity to unmeasured bias through study design. Epidemiology **7**, 79–87 (2013)

Chapter 4
Competing Theories Structure Design

Abstract In a well-designed experiment or observational study, competing theories make conflicting predictions. Several examples, some quite old, are used to illustrate. Also discussed are: the goals of replication, empirical studies of reasons for effects, and the importance of systemic knowledge in eliminating errors.

> About thirty years ago there was much talk that geologists ought only to observe and not theorise; and I well remember some one saying that at this rate a man might as well go into a gravel-pit and count the pebbles and describe the colours. How odd it is that anyone should not see that all observation must be for or against some view if it is to be of any service.
>
> Charles Darwin [17]
> Letter to Henry Fawcett

> What goes on in science is not that we try to have theories that accommodate our experiences; it's closer that we try to have experiences that adjudicate among our theories.
>
> Jerry Fodor [22, pages 202–203]

> It takes a theory to kill a theory … given our need to have a systematic way of thinking about complicated reality.
>
> Paul A. Samuelson [58, page 304]

4.1 How Stones Fall

In his *Physics*, Aristotle claimed a heavy object falls faster than a light one. Much in daily experience confirms this, or seems to. Stones fall faster than feathers, for instance. Everyone has seen this. How can you doubt what you and everyone else have seen?

© Springer Nature Switzerland AG 2020

P. R. Rosenbaum, *Design of Observational Studies*, Springer Series in Statistics,
https://doi.org/10.1007/978-3-030-46405-9_4

Galileo doubted that Aristotle had it right. In his dialogue[1] *Two New Sciences* [26], Galileo proposed a thought experiment, perhaps the most famous of all thought experiments. Suppose Aristotle were correct, says Galileo, and suppose we connected a large stone to a small one. Would the two connected stones fall faster or slower than the large stone falling alone? In Galileo's words [26, pages 66–67]:

> *Salviati*: But without other experiences, by a short and conclusive demonstration, we can prove clearly that it is not true that a heavier moveable is moved more swiftly than another, less heavy, these being of the same material, and in a word, those of which Aristotle speaks. ... if we had two moveables whose natural speeds were unequal, it is evident that were we to connect the slower to the faster, the latter would be partly retarded by the slower, and this would be partly speeded up by the faster. Do you not agree with me in this opinion?
>
> *Simplicio*: It seems to me that this would undoubtedly follow.
>
> *Salviati*: But if this is so, and it is also true that a large stone is moved with eight degrees of speed, for example, and a smaller one with four degrees, then joining both together, their composite will be moved with a speed less than eight degrees. But the two stones joined together make a larger stone than that first one which was moved with eight degrees of speed; therefore this greater stone is moved less swiftly than the lesser one. But this is contrary to your assumption. So you see how, from the supposition that the heavier body is moved more swiftly than the less heavy, I concluded that the heavier moves less swiftly.
>
> *Simplicio*: I find myself in a tangle ...

Galileo develops his own theory involving a "law of uniformly accelerated motion": [26, page 166]:

> Proposition II. Theorem II: If a moveable descends from rest in uniformly accelerated motion, the spaces run through in any times whatever are ... as the squares of those times.

In later Newtonian terms, if an object is acted upon by a single, constant force, namely gravity, its acceleration will be constant in time, its velocity will increase linearly with time, and the distance it travels will increase with the square of the time spent traveling.[2] Galileo's proposition concerns a theory that asserts that the instantaneous velocity of a falling stone is ever increasing at a constant rate of increase, is never the same for two distinct instants. Instantaneous velocity is not measurable, but the proposition says the theory has a measurable consequence, a relationship between distance traveled and time spent traveling. Naturally, Galileo set up an experiment to test the testable consequence of his theory. Or, at least, it seems natural to us, though it was something of a new idea at the time. Things fall

[1]The full title is "Discourses and Mathematical Demonstrations Concerning Two New Sciences Pertaining to Mechanics and Local Motions" by Galileo Galilei, chief philosopher and mathematician to the most serene Grand Duke of Tuscany. It was written while Galileo was under house arrest for teaching the Copernican theory. Galileo could not publish it in Italy, and it was published by Elzevir in Leyden in 1638.

[2]This is, of course, a simple calculus problem. Then again, calculus was invented, in part, with a view to making this a simple problem.

quickly, making measurement difficult, so Galileo began by slowing the speed of the fall. Harré [30, pages 79–81] writes:

> The experiment involved cutting and polishing a groove in a wooden beam and lining the groove with parchment. A polished bronze ball was let roll down the groove when the beam was set on an incline … Variations in time for many runs of the same descent were very small. The theoretically derived relation between distances and times for uniformly accelerating motion was tested by letting the ball roll a quarter, then half, then two-thirds and so on, of the length of the groove, measuring the times for the journey in each case. The ball did indeed take half the time required for a full descent to reach the quarter way point.

Much is instructive in Galileo's approach.

- Galileo develops his theory in dialogue with, in opposition to, existing theories.[3]
- Everyday impressions that might casually be taken to support Aristotle's theory are immediately challenged, not as false impressions, but as providing no support to Aristotle's theory. Galileo turns attention to a "heavier moveable … of the same material." If stones fall faster than feathers, but heavy, larger stones fall no faster than lighter, smaller stones of the same material, then that is evidence against, not evidence in favor of Aristotle's theory, for it says that something besides weight causes stones and feathers to fall at different rates.[4] Repeatedly, either by abstraction in argument or by experimental procedure—"polishing a groove … lining the groove with parchment … a polished bronze ball"—disturbing influences are removed so that weight and weight alone varies.[5]
- The case Galileo makes is neither strictly theoretical nor strictly empirical. Using a thought experiment, he argues that Aristotle's theory contradicts itself, that it is necessarily false without reference to any particular experimental observation. The thought experiment is not totally convincing: if an actual experiment failed to reproduce the thought experiment, we would be puzzled rather than certain that the actual experiment erred. Nonetheless, Galileo's theoretical argument creates a space for competing theories and their experimental evaluation.[6]

---

[3]The importance of working with several theories at once is often stressed. Paul Feyerabend writes: "You can be a good empiricist only if you are prepared to work with many alternative theories rather than with a single point of view and 'experience.' … Theoretical pluralism is assumed to be an essential feature of all knowledge that claims to be objective … [20, pages 14–15] [T]he evidence that might refute a theory can often be unearthed only with the help of an incompatible alternative [21, page 29]." See also [12, 48].

[4]Robert Nozick [45, pages 261–263], Peter Achinstein [1], and Kent Staley [62] argue that whether or not E constitutes evidence for T is a matter itself open to empirical challenge and investigation.

[5]The systematic exclusion of sources of variation besides the one cause under study is familiar in every scientific laboratory, and it was named the "method of difference" by John Stuart Mill [40]; see Chap. 16.

[6]Thomas Kuhn [35] and J.R. Brown [5] present perspectives on the role of thought experiments in scientific work. In particular, Kuhn writes [35, page 264]: "By transforming felt anomaly to concrete contradiction, the thought experiment [… provided the…] first clear view of the misfit between experience and implicit expectation …".

- Galileo introduces a competing theory, a beautiful theory, in terms of constant acceleration. In point of fact, this theory speaks about things that cannot be seen or measured, namely instantaneous velocity and how it changes. Galileo develops an observable, testable consequence of this theory. The testable consequence involves very precise predictions concerning the relationship between distance traveled and time spent traveling.
- The experiment is limited in scope and is atypical of situations in which objects fall. The descent is slowed by a beam set on an incline, permitting precise measurements to be compared with a theory that makes precise predictions. The object, a polished bronze ball, is atypical of falling objects. And so on. No attempt is made to survey the falls of all of the world's falling objects, because those comparisons would be subject to innumerable disturbing influences that would obscure the issue under examination. Laura Fermi and Gilberto Bernardini [19, page 20] write that Galileo's experiments "reproduced the essential elements of the phenomenon under controlled and simplified conditions."
- Galileo conducted a severe test of his theory, but did little to demonstrate that it is always true: perhaps the theory is correct only for one bronze ball, on one inclined beam, with suitable parchment, in Italy, in the 1600s.[7]

4.2 The Permanent-Debt Hypothesis

In his permanent income hypothesis, Milton Friedman argued that personal consumption today is guided by expected long-term income, not current income or cash on hand.[8] Much in daily experience confirms this, or seems to. At a business school, doctoral students receive tuition and a stipend but spend modestly, while MBAs pay tuition with no stipend, yet spend less modestly. This is consistent with the permanent income hypothesis: MBAs expect to earn more in the long run. Whatever may be the ultimate fate of the permanent income hypothesis, it embodies a plausible claim: a rational person would anticipate future income in deciding about current consumption.[9]

[7]The notion that scientific theories are testable but not demonstrable is thematic in the work of Sir Karl Popper [51]. He writes: "Theories are not verifiable, but they can be 'corroborated' … [W]e should try to assess what tests, what trials, [the theory] has withstood [51, page 251] … [I]t is not so much the number of corroborating instances which determines the degree of corroboration as the severity of the various tests to which hypothesis can be, and has been, subjected [51, page 267]." Obviously, if you doubted Galileo, you could polish a bronze ball ….

[8]Friedman discusses his permanent income hypothesis in his *A Theory of the Consumption Function* [24]. A concise, simplified version of the permanent income hypothesis is given in a couple of pages by Romer [54, Chapter 7]. Zellner [76, III.C] and Friedman [25, Chapter 12] reprint sections of Friedman's *A Theory of the Consumption Function*.

[9]Perhaps there is a certain similarity between a rational person and a polished bronze ball: both are uncommon in nature, but their behavior is of interest nonetheless.

The permanent income hypothesis has been challenged in various ways. David Gross and Nicholas Souleles [28, page 149] write:

> The canonical Permanent-Income Hypothesis (PIH) assumes that consumers have certainty-equivalent preferences and do not face any liquidity constraints. Under these assumptions the marginal propensity to consume (MPC) out of liquid wealth depends on model parameters, but generally averages less than 0.1. The MPC out of predictable income or "liquidity" (e.g., increases in credit limits), which do not entail wealth effects, should be zero. The leading alternative view of the world is that liquidity constraints are pervasive. Even when they do not currently bind, they can be reinforced by precautionary motives concerning the possibility that they bind in the future. Under this view the MPC out of liquidity can equal one over a range of levels for "cash-on-hand," defined to include available credit.

As in Sect. 4.1, the starting point is not one theory, but the contrast between two theories. Contrast must become conflict: one must find a quiet undisturbed location where contrasting theories make conflicting predictions. Gross and Souleles [28, pages 150–151] continue:

> To test whether liquidity constraints and interest rates really matter in practice, this paper uses a unique new data set containing a panel of thousands of individual credit card accounts from several different card issuers. The data set ... includes essentially everything that the issuers know about their accounts, including information from people's credit applications, monthly statements, and credit bureau reports. In particular, it separately records credit limits and credit balances, allowing us to distinguish credit supply and demand, as well as account specific interest rates. These data allow us to analyze the response of debt to changes in credit limits and thereby estimate the MPC out of liquidity, both on average and across different types of consumers. The analysis generates clean tests distinguishing the PIH, liquidity constraints, precautionary saving, and behavioral models of consumption ...

The permanent income hypothesis predicts that an offer of credit unrelated to permanent income should not prompt an increase in spending. Gross and Souleles are seeking a circumstance in which there is a change in available credit with no change in expected long-term income. There is a great deal of activity in these credit card accounts that obscures the intended contrast of theories, and efforts are needed to remove these irrelevant disturbances. Gross and Souleles adopt two strategies: "first, we use an unusually rich set of control variables..." [28, page 154]. For instance, a person might be offered more credit because of increased credit scores or debts that appear to be under control, and perhaps these are related to income, but Gross and Souleles know what the card issuers know about these matters, so variations in these quantities can be controlled. Second, they exploit "timing rules" built into the credit supply functions ... many issuers will not consider ... an account for a line change if it has been less than six months or less than one year since the last line change" [28, page 155]. Their point is that these timing rules generate small jumps in available credit that are not much related to an individual's financial circumstances: if your credit account is entirely unchanged in all its aspects (or more accurately, if this is the case after adjustment for the many control variables), you will be offered more credit after a lapse of time. Gross and Souleles are trying, perhaps successfully, to isolate a change in credit availability accompanied by no other consequential change. They conclude [28, page 181]: "We

found that increases in credit limits generate an immediate and significant rise in debt, counter to the PIH."

The goal here is no different than the goal in Sect. 4.1: to display an effect that contrasts two theories as sharply as possible, removing to the greatest degree possible all disturbances that might obscure that contrast. There is no attempt to generalize from a sample to a population. The two theories began as general theories; there is no need to generalize further. The goal is to contrast conflicting predictions of two general theories in the stillness of a laboratory.

4.3 Guns and Misdemeanors

The 1968 Federal Gun Control Act prohibits the purchase of guns by felons. Beginning in 1991, California prohibited the purchase of handguns by individuals convicted of violent misdemeanors, including assault or brandishing a firearm. It is not clear whether such prohibitions are effective at reducing gun-related violence. On one theory, it is fairly easy to purchase a handgun illegally, for instance, by purchasing the gun from someone who can buy it legally, so the prohibitions deter only those who wish to avoid illegal activity, perhaps not the best target for deterrence. On a different theory, legal prohibitions backed by punishment deter individuals who wish to avoid punishment.

The two theories are plain enough, but contrasting them is not so easy. Under what circumstances do the two theories make conflicting predictions? Under both theories, it would not be surprising if violent men remain violent, with or without restrictions on handgun purchases. The laws restrict purchases by violent individuals; these individuals cannot be compared with unrestricted, nonviolent individuals.

The effects of California's law were investigated by Garen Wintemute, Mona Wright, Christiana Drake, and James Beaumont [74] in the following way. They compared two groups of Californians who had been convicted of violent misdemeanors of the sort that would have prevented a legal purchase of a handgun beginning in 1991. One group consisted of individuals who had applied to purchase a handgun in 1989 or 1990, before the law took effect. The second group consisted of individuals who applied to purchase a handgun in 1991 and whose application was denied. These two groups may not be perfectly comparable, but at least both groups had been convicted of violent misdemeanors and both groups sought to purchase a handgun. In terms of demographics and previous convictions, the two groups looked fairly similar [74, Table 1]. If California's law were effective, one expects to see a lower rate of gun and violent crime in the group denied a handgun purchase, with no difference in nongun or nonviolent crime. If the law were ineffective, one expects to see similar rates of gun and violent crime in both groups. With or without adjustments for demographics and previous convictions, Wintemute et al. [74, Tables 2 and 3] found lower rates of gun and violent crime

in the group denied a handgun purchase, with little indication of a difference in the rate of other crimes.

The reasoning here is much the same as in Sects. 4.1 and 4.2. Admittedly, the two theories are quite plain: one says the policy works, the other denies this. And yet, because the policy is targeted at offenders, care is needed to identify circumstances in which these theories make conflicting predictions.

4.4 The Dutch Famine of 1944–1945

Does malnutrition in utero reduce mental performance at age 19? The two theories are straightforward: it does or it does not. It is not straightforward, however, to identify circumstances in which these theories lead to different predictions. Reliable information is needed about mental performance at age 19 and about mother's diet 20 years prior to that. Controls are needed whose mothers were not malnourished but who were otherwise similar in background, education, social class, etc. Neither such information nor such controls are commonly available.

Zena Stein, Mervyn Susser, Gerhart Saenger, and Francis Marolla [64] found what was needed in the Dutch famine of 1944–1945. They wrote [64, page 708]:

> On 17 September 1944 British paratroops landed at Arnhem in an effort to force a bridgehead across the Rhine. At the same time, in response to a call from the Dutch government-in-exile in London, Dutch rail workers went on strike. The effort to take the bridgehead failed, and the Nazis in reprisal imposed a transport embargo on western Holland. A severe winter froze the barges in the canals, and soon no food was reaching the large cities ... At their lowest point the official food rations reached 450 calories per day, a quarter of the minimum standard. In cities outside the famine area, rations almost never fell below 1300 calories per day. ... The Dutch famine was remarkable in three respects: (i) Famine has seldom if ever struck where extensive, reliable and valid data allow the effects to be analyzed within specified conditions of social environment. (ii) The famine was sharply circumscribed in both time and place. (iii) The type and degree of nutritional deprivation during the famine were known with a precision unequaled in any large human population before or since.

To know a child's date and place of birth is to know whether the mother was affected by the famine. So Stein et al. defined famine and control regions in Holland, as well as a cohort of children born just before the start of the famine, several cohorts exposed to the famine at various times in utero, and a cohort conceived after the famine ended. In these cohorts, virtually all males received medical examinations and psychological testing, including an IQ test, in connection with induction into the military. For 1700 births at two hospitals, one in a famine region (Rotterdam) and one in a control region (Heerlan), birth weights were available. Stein et al. [64, page 712] concluded: "Starvation during pregnancy had no detectable effects on the mental performance of surviving male offspring."

Subsequent studies used the Dutch famine to examine other outcomes, including schizophrenia, affective disorders, obesity, and breast cancer; see [37, 66] for reviews of these studies. For studies that have taken a similar approach in other situations, see [3, 60].

The Dutch famine was unrepresentative both of typical in utero development and typical famines, and it was useful for studies of development precisely in the ways it was unrepresentative. In representative situations, famine is confounded with other factors that obscure the effects caused by malnutrition in utero.

4.5 Replicating Effects and Biases

The randomized experiment in its idealized, perhaps unattainable, form is subject to a single source of uncertainty—that stemming from a finite sample size. In this ideal, biases of every sort have been eliminated by blocking and randomization; therefore, the sole function of replication is to increase the sample size. Actual experiments are a step or two removed from the ideal, and observational studies are several steps removed from that. In an observational study, even an excellent one, there is the possibility that can never be entirely eliminated: treated and control subjects who look comparable may not actually be comparable, so differing outcomes may not be treatment effects.[10] In observational studies, the principal and important function of replication is to shed some light on biases of this sort.

Here, too, there are two theories. The first theory asserts that previous studies have produced estimates of treatment effects without much bias. The competing theory denies this, asserting instead that previous studies were biased in certain particular ways, and if those biases were removed, then the ostensible treatment effect would be removed with them. The replication does not repeat the original study; rather, it studies the same treatment effect in a context that is not subject to particular biases claimed by the competing theory. Replication is not repetition: the issue at hand is whether the ostensible effect can be reproduced if a conjectured bias is removed.

An example is David Card and Alan Krueger's [9, 10] replication of their earlier study [8] of the effects of the minimum wage on employment[11]; see also Sect. 12.3. On 1 April 1992, New Jersey raised its minimum wage by about 20%, whereas neighboring Pennsylvania left its minimum wage unchanged. In their initial study, Card and Krueger [8] used survey data to examine changes in employment at fast-food restaurants, such as Burger King and Wendy's, in New Jersey and eastern Pennsylvania, before and after the increase in New Jersey's minimum wage. They found "no evidence that the rise in New Jersey's minimum wage reduced

[10] In an observational study, an increase in sample size ensures only that the estimator is closer to whatever the estimator estimates. Typically, the estimator estimates the sum of a treatment effect and a bias whose unknown magnitude does not decrease as the sample size increases. When a bias of this sort is not zero, as the sample size increases, confidence intervals shrink in length to exclude the true treatment effect, and hypothesis tests are ever more likely to reject the true hypothesis. It is a mistake to attach great importance to an increase in the sample size (or to an increase in statistical efficiency) when substantial biases of fixed size are present or likely.

[11] For a brief survey of the economics of the minimum wage, see [6, §12.1].

employment at fast-food restaurants in the state." Their careful and interesting study received some critical commentary; e.g., [44]. One of several issues was the quality of the employment data obtained by a telephone survey, and another was simply that New Jersey and Pennsylvania differ in many ways relevant to employment change, not just in their approach to the minimum wage. In 1996, the minimum wage as set by the U.S. Federal Government was raised, forcing up the minimum wage in Pennsylvania, without forcing changes in New Jersey whose state minimum wage was already above the new Federal minimum wage. Card and Krueger [9, 10] then replicated the original study with the roles of New Jersey and Pennsylvania now reversed, using employment data from payroll records for Unemployment Insurance made available by the U.S. Bureau of Labor Statistics. Their findings were similar. The replication does not eliminate all concerns raised about the original study, but it does make two specific concerns less plausible.

Mervyn Susser writes:

> The epidemiologist [...seeks...] consistency of results in a variety of repeated tests ... Consistency is present if the result is not dislodged in the face of diversity in times, places, circumstances, and people, as well as of research design [68, page 88] ... The strength of the argument rests on the fact that diverse approaches produce similar results [67, page 148].

Susser's careful statement is easily, perhaps typically, misread. Mere consistency upon replication is not the goal. Mere consistency means very little. Rather, the goal is "consistency ... in a variety of ... tests" and "consistency ... not dislodged in the face of diversity." The mere reappearance of an association between treatment and response does not convince us that the association is causal—whatever produced the association before has produced it again. It is the tenacity of the association—its ability to resist determined challenges—that is ultimately convincing. For a similar view with examples from physics, see Allan Franklin and Colin Howson [23]. For a similar view with examples from mathematics in a Bayesian formulation, see Georg Polya [50, especially pages 463–464].

To illustrate, consider two sequences of studies, one sequence involving the effectiveness of treatments for drug addiction, the other involving the effects of advertising on prices. The first sequence increases the sample size, but the bias of greatest concern remains the same. The second sequence asks the same question in very different contexts.

Several nonexperimental studies of the effects of treatments for heroin or cocaine addiction have found that people who remain in treatment for at least three months are more likely to remain drug-free than people who drop out of treatment before three months. These studies had involved the Drug Abuse Reporting Program (DARP) and the Treatment Outcomes Prospective Study (TOPS). In a third study using a different source of data, Hubbard et al. [32, page 268] wrote:

> The general finding from DARP and TOPS that treatment duration of at least three months is associated statistically and clinically with more positive outcomes supports the inference of treatment effectiveness. The following analysis retests this hypothesis ...

In their new study, Hubbard et al. [32] found, again, that people who dropped out of treatment went on to use more illegal drugs than people who remained in treatment. Is this an effect caused by treatment? Or is it simply that a person who has little enthusiasm for ending his addiction is both more likely to drop out of treatment and more likely to use illegal drugs? A summary report of the National Academy of Sciences [38, page 17] was understandably skeptical:

> Depending on the process by which users are selected into treatment programs and the determinants of dropout from such programs … the data may make treatment programs seem more or less cost-effective than they actually are.

These three studies are all open to the same question: Is a comparison of people who remain in treatment and people who drop out revealing an effect caused by the treatment, or is it revealing something about the people who stay in treatment and those who drop out? Because the same question can be asked of each study, the replications increase the sample size—which was large from the start, and never much at issue—but they do not make progress towards answering a basic concern.

What is the effect of advertising on prices? More precisely, what is the effect on prices of imposing or removing a restriction on the advertising of prices? One study by Amihai Glazer [27] made use of a strike from 10 August 1978 to 5 October 1978 that shut down three daily New York City newspapers, the New York Times, the New York Post, and the Daily News. The strike reduced advertising of retail food prices in New York City. Glazer looked at changes in retail food prices in Queens, the eastern borough of New York City, comparing Queens with adjacent Nassau County, outside the City, where the major newspaper, Newsday, was unaffected by the strike. A second study by Jeffrey Milyo and Joel Waldfogel [41] examined changes in liquor prices in Rhode Island before and after the U.S. Supreme Court struck down Rhode Island's ban on liquor advertising, comparing Rhode Island with adjacent areas of Massachusetts that were unaffected by the Court decision. In a third study, C. Robert Clark [13] examined the effect on the prices of children's breakfast cereals of Quebec's ban on advertising aimed at children under the age of 13. Clark compared the prices of cereals for children and cereals for adults in Quebec and in the rest of Canada. Each situation—the newspaper strike in New York, the Supreme Court decision in Rhode Island, the ban on advertising to children in Quebec—has idiosyncrasies that might be mistaken for an effect of advertising on prices, but there is no obvious reason why these idiosyncrasies should align to reproduce the same association between advertising and prices. Unlike the studies of treatments for drug addiction, where one alternative explanation suffices for three studies, in the advertising studies, three unrelated idiosyncrasies would have to produce similar associations. That could happen, but it seems less and less plausible as more and more studies are designed to replicate an effect, if it is a real effect, but to avoid replicating biases.

The goal in replication is to study the same treatment effect in a new context that removes some plausible bias that may have affected previous studies [36, 56].

4.6 Reasons for Effects

> To suppose universal laws of nature capable of being apprehended by the mind and yet having no reason for their special forms, but standing inexplicable and irrational, is hardly a justifiable position. Uniformities are precisely the sort of facts that need to be accounted for. . . . Law is *par excellence* the thing that wants a reason.
>
> Charles Sanders Peirce [47]
> The Architecture of Theories

There is an obvious distinction between empirical evidence in support of a theory and a reasoned argument in support of a theory, but as is often the case with obvious distinctions, this one becomes less obvious upon close inspection. It is, after all, common to argue that what was taken as empirical evidence in support of a theory is not evidence at all.[12] It is equally common to cast doubt upon a reasoned argument with the aid of empirical evidence. To offer a reasoned argument in support of the claim that a treatment does or should or would have a certain effect is to create a new object for empirical investigation, namely the argument itself.[13] Consider an example.

In an effort to reduce gun violence, various municipalities have created programs to buy handguns from their owners. For instance, Milwaukee did this in 1994–1996, offering $50 for a workable gun. Buyback programs are popular because they are voluntary: they do not meet the same resistance that a coercive program might meet. However, voluntary programs may also be less effective than coercive programs. Do voluntary buyback programs work to reduce gun violence? One could argue that buybacks have to work: whenever a gun is destroyed, that gun cannot be used in a future homicide or suicide. One could argue that the program cannot really work: a rational person would turn in a gun for $50 only if he does not plan to use it, so the program is buying guns that would otherwise rest harmlessly in the attic. One could argue that buybacks work because people are not rational: many gun deaths are accidents, and others stem from impulsive responses to rage or depression, so buying guns from people who do not plan to use them helps people to adhere to their peaceful plans. (See Ainslie [2] or Rachlin [53] for cogent general discussions of rational control of irrational impulses.) One could argue that buybacks cannot really work because many violent individuals do not have peaceful plans and will not sell their guns, intending to thwart their own violent inclinations: "[T]he self-images of

[12] See Sect. 4.1 and Note 4.

[13] Dretske [18, pages 20–22] writes: "More often than not a reason, or the giving of reasons, supplies a recipe . . . for the possible falsification of the statement or statements for which the reason is given. By this I mean that although some statements, considered in isolation, may appear irrefutable, they lose this invulnerability when taken in the context provided by evidential support. . . . If Q is a reason for believing P true, then although P need not be true, Q must be. We can, and often do, give reasons—sometimes very good reasons—for believing something that is not true . . . [A]ny challenge to the truth of Q is simultaneously a challenge to the acceptability of Q as a reason."

violent criminals are always congruent with their violent criminal actions," wrote Athens [4, page 68] in a careful, perceptive study. In short, it is not difficult to offer a variety of reasons in support of a variety of possible effects. Whatever the strengths or weakness of these arguments as arguments for or against particular effects, the arguments create objects for empirical investigation, namely the validity of the arguments themselves.

Basic to each of these arguments is the question: Do buyback programs buy the guns that would otherwise be used in gun violence? One aspect of this question is: Do buyback programs typically buy the types of guns typically used in gun violence? A second aspect is: Do buyback programs buy guns in sufficient quantity to affect gun violence? Evelyn Kuhn and colleagues [34] addressed the first aspect and a report of the U.S. National Academy of Sciences [72, pages 95–96] addressed the second.

Table 4.1 is Kuhn et al.'s [34] comparison of the caliber of repurchased guns and homicide guns. To a dramatic degree, with an odds ratio of ten or more, repurchased guns are small caliber and homicide guns are large caliber. Kuhn et al. make several similar comparisons, finding that repurchased guns tended to be small, obsolete revolvers, whereas homicide guns tended to be large, inexpensive, semiautomatic pistols. They found similar but not identical results for suicide guns. The buyback programs are buying types of guns infrequently used in gun violence.

The report of the U.S. National Academy of Sciences makes this same point and two others [72, pages 95–96]:

> The theory on which gun buy-back programs is based is flawed in three respects. First, the guns that are typically surrendered in gun buy-backs are those that are least likely to be used in criminal activities … Second, because replacement guns are relatively easily obtained, the actual decline in the number of guns on the street may be smaller than the number of guns that are turned in. Third, the likelihood that any particular gun will be used in a crime in a given year is low. In 1999, approximately 6,500 homicides were committed with handguns. There are approximately 70 million handguns in the United States. Thus, if a different handgun were used in each homicide, the likelihood that a particular handgun would be used to kill an individual in a particular year is 1 in 10,000. The typical buy-back program yields less than 1,000 guns.

In brief, their third point claims that the scale of buyback programs is too small to have a meaningful effect.

Table 4.1, related results in [34], and the arguments in the quoted paragraph from [72, pages 95–96] are intended to undermine the reasoning that justifies gun buyback programs, but they provide no direct evidence about the effects of such programs on violence. In [34], violence is not measured. Table 4.1 and the quoted paragraph

Table 4.1 Homicide and buyback handguns in Milwaukee by caliber. As the odds ratios indicate, homicide and buyback guns are quite different

Caliber	Buyback	Homicide	Odds Ratio
Small: .22, .25, .32	719	75	1.0
Medium: .357, .38, 9mm	182	202	10.6
Large: .40, .44, .45	20	40	19.2
Total	941	369	

attack the idea that gun buyback programs could work by buying the guns used in gun violence, because the programs buy the wrong types of guns and too few of them. It is, nonetheless, not entirely inconceivable that gun buyback programs do work to reduce violence, but do this in some other way. Imagine a highly publicized buyback program, in which, each night, the television news interviews someone who turned in a gun, reporting the accumulated total of guns repurchased, accompanied by comments from civic leaders about how the city is "getting rid of its guns and ridding itself of violence." Even if the program bought the wrong guns and too few of them, it is not entirely inconceivable that the publicity, fanciful though it may be, would affect public sentiment in a way that ultimately affected the level of gun violence. Of course, in arguing for a new policy, one might hope to say more than: "it is not entirely inconceivable that the policy might work." In the discussion here, a precise distinction is intended between "direct evidence of a treatment effect" and a "reason for a treatment effect." The distinction does not refer to the presence or absence of empirical evidence: empirical evidence is present in some form in most scientific work. Nor does the distinction refer to the strength or quality of the evidence: in the case of gun buybacks, the evidence against the reasons for an effect may be more compelling than the direct evidence of ineffectiveness [72, pages 95–96]. "Direct evidence of a treatment effect" means a study of the relevant treatment, on the relevant subjects or units, measured in terms of the relevant outcome. A randomized experiment or an observational study may provide direct evidence of an effect in this sense, and the evidence is likely to be stronger if it comes from a well-conducted randomized experiment. In contrast, an empirical study of a "reason for an effect" is a study that is a step removed; it provides a reason for thinking the relevant treatment would have a particular effect were it to be applied to the relevant subjects measured in terms of the relevant response, but one or more of these elements is not actually present. In the case of gun buybacks in Table 4.1 and in [34], the relevant outcome is the level of gun violence, which is not measured; instead, the results refer to the types of guns that are repurchased. In the paragraph above quoted from [72, pages 95–96], the relevant outcome is again gun violence, but the results refer to the number of guns repurchased. Dafna Kanny and colleagues [33] studied the effects on helmet use of a law requiring bicyclists to wear helmets: this provides a plausible reason for thinking the law reduced accidents, but it is a step removed, because helmet use rather than accident reduction is the outcome. In the 1964 U.S. Surgeon General's Report, *Smoking and Health*, it is noted that [63, page 143] "there is evidence from numerous laboratories that tobacco smoke condensates and extracts of tobacco are carcinogenic for several animal species." This is a reason for thinking that tobacco is carcinogenic in humans, but it is a step removed, because the subjects are not human. Much laboratory work in the biomedical sciences studies reasons for effects in humans by examining effects on animals or cell cultures or molecules. The use of surrogate outcomes in medicine is the study of a reason for an effect in the sense defined here; for example, the use of colonic polyps as a surrogate for mortality from colorectal cancer [59]. Laboratory style experimentation in economics studies reasons for effects when small experimental incentives are proxies for incentives faced in actual economic

decisions [49]. A theoretical argument about what effect would appear in a tidy and simple theoretical world may constitute a reason for an effect in the untidy and complex world we actually inhabit; see, for instance, Sect. 12.3.

There is no substitute for direct evidence of the effect of a treatment. And yet, evidence about reasons for an effect remains important. Evidence about reasons for an effect may either strengthen or undermine direct evidence of an effect, and either outcome is constructive [57, 70].

4.7 The Drive for System

> A good scientific theory is under tension from two opposing forces: the drive for evidence and the drive for system. Theoretical terms should be subject to observable criteria, the more the better; and they should lend themselves to systematic laws, the simpler the better, other things being equal. If either of these drives were unchecked by the other, it would issue in something unworthy of the name of scientific theory: in the one case a mere record of observations, and in the other a myth without foundation.

W.V.O. Quine [52, page 90]

> [I]mpersonal knowledge … has a special commitment to reasons. Bodies of knowledge are essentially, if to varying degrees with different subjects, systematic. There is both a pure and an applied reason for this. Pure, because the aim is not just to know but to understand, and in scientific cases at least understanding necessarily implies organization and economy. Applied, because a body of knowledge will only be freely extensible and open to criticism if rationally organized. … So knowledge, in this sense, must have reasons.

Bernard Williams [73, page 56]

The first paragraph of Chap. 1 recalled Borgman's political cartoon in which a TV newsman presents "Today's Random Medical News" that "according to a report released today … coffee can cause depression in twins." The cartoon accurately depicts a popular misconception. It is often thought, incorrectly, that a scientific study has a clear and stable interpretation on the day it is published, and that interpretation is "news." In fact, the considered judgment about a scientific study hardens slowly, like cement, as the study is subjected to critical commentary, reconciled with other studies, both past and future, and integrated into a systemic understanding of the larger topic to which the study is a contribution. The science of the undergraduate classroom—that is, the science that is presented simply as the way things are—is not a list of findings from a sequence of individually decisive studies, but rather the systemic reconciliation and integration of many, perhaps conflicting, studies. The study of "reasons for effects" in Sect. 4.6 is part of this systemic reconciliation. Gaps and conflicts in systemic knowledge provoke new investigations aimed at closing gaps, resolving conflicts. In this sense, there is no "science news." On the day of its release, the new report faces a fate that is far too uncertain to constitute news. By the time a systemic understanding has hardened, by the time today's report has found a resting place in that understanding, both the report and the systemic understanding are far too old to constitute news.

Consider, for instance, Steven Clinton and Edward Giovannucci's [14] review in 1998 of the possibility that diet affects the risk of prostate cancer. Their review attempted to organize, reconcile, and lend systematic form to 193 scientific reports. The following excerpts indicate the flavor of their review.

> Studies of prostate cancer have yielded inconsistent results concerning the relationship between various measurements of adult body mass or obesity and risk [of prostate cancer, citing 11 studies, which are then contrasted in detail] [14, page 422]. ... An association between prostate cancer incidence and tobacco use has been inconsistent ... Additional studies focusing on the critical timing of smoking during the life cycle ... are necessary [14, page 419]. ... The relationship between alcohol consumption and risk of prostate cancer has been evaluated in a series of studies. No strong evidence has emerged for an association [citing 4 studies] [14, page 424]. ... Reports of a correlation between diets rich in meat or dairy products and risk of prostate cancer are frequent [citing 12 studies] [14, page 425]. ... [Citing Giovannucci's own study, in which a nutrient, lycopene, found in tomatoes, was studied:] men in the highest quintile of lycopene intake experienced a 21% lower risk [... but ...] investigators should use caution in assuming that lycopene mediates a relationship between consumption of tomato products and lower risk of prostate cancer [14, page 429]. ... A recent study was designed to examine the effects of selenium supplementation on recurrence of skin cancer in a high-risk population [citing a randomized clinical trial]. Selenium treatment did not influence the risk of skin cancer, although selenium-treated patients had a nonsignificant reduction in total cancer mortality, as well as a lower risk of prostate cancer. Additional studies are necessary ... [14, pages 429–30]

The serious, evenhanded, cautious tone of Clinton and Giovannucci's review is in marked contrast to the presentation of diet and health in the popular press. Clinton and Giovannucci's calls for additional studies are often cited by subsequent studies [42, 46]. An updated review six years later reports some progress but many questions still open [39].

Most scientific fields seek a systemic understanding that: (i) contrasts and reconciles past studies, (ii) guides future studies towards promising gaps in the current understanding, and (iii) integrates theoretical and empirical work. For an example from economics, see Jonathan Gruber's [29] review of the economic issues related to providing health insurance to the uninsured. For an example from criminology, see the report by Charles Wellford and colleagues [72] concerning firearms and violence. For an example from the epidemiology of dementia emphasizing methodology as the origin for inconsistent results, see the review by Nicola Coley, Sandrine Andrieu, Virginie Gardette, Sophie Gillette-Guyonnet, Caroline Sanz, Bruno Vellas, and Alain Grand [15].

The drive for system plays a key role in eliminating scientific error, but it also plays a key role in deciding what to study next. A new experiment or observational study contributes to an extant understanding, one that is partly systematic, partly incomplete, partly conflicting, and doubtless partly, perhaps substantially, in error. A new study is judged by its contribution to organizing, completing, reconciling, or correcting that understanding.

4.8 Further Reading

Karl Popper [51], Georg Polya [50], John Platt [48], and Paul Feyerabend [20] offer abstract discussions relevant to the current chapter; see also [55]. The material in Sects. 4.5 and 4.6 is discussed in greater detail in [36, 56] and [57], respectively. Replication is further discussed in [36, 43].

Bibliography

1. Achinstein, P.: Are empirical evidence claims a priori? Br. J. Philos. Sci. **46**, 447–473 (1995)
2. Ainslie, G.: Breakdown of Will. Cambridge University Press, New York (2002)
3. Almond, D.: Is the 1918 influenza pandemic over? Long-term effects of in utero influenza exposure in the post-1940 US population. J. Polit. Econ. **114**, 672–712 (2006)
4. Athens, L.: Violent Criminal Acts and Actors Revisited. University of Illinois Press, Urbana (1997)
5. Brown, J.R.: The Laboratory of the Mind: Thought Experiments in the Natural Sciences. Routledge, New York (1991)
6. Cahuc, P., Zylberberg, A.: Labor Economics. MIT Press, Cambridge, MA (2004)
7. Card, D.: The causal effect of education. In: Ashenfelter, O., Card, D. (eds.) Handbook of Labor Economics. North-Holland, New York (2001)
8. Card, D., Krueger, A.: Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania. Am. Econ. Rev. **84**, 772–793 (1994)
9. Card, D., Krueger, A.: A reanalysis of the effect of the New Jersey minimum wage increase on the fast-food industry with representative payroll data. NBER Working Paper 6386. National Bureau of Economic Research, Boston (1998)
10. Card, D., Krueger, A.: Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania: Reply. Am. Econ. Rev. **90**, 1397–1420 (2000)
11. Card, D., DelloVigna, S., Malmendier, U.: The role of theory in field experiments. J. Econ. Perspect. **25**, 39–62 (2011)
12. Chamberlin, T.C.: The method of multiple working hypotheses. Science **15**, 92 (1890). Reprinted Science **148**, 754–759 (1965)
13. Clark, C.R.: Advertising restrictions and competition in the children's breakfast cereal industry. J. Law Econ. **50**, 757–780 (2007)
14. Clinton, S.K., Giovannucci, E.: Diet, nutrition, and prostate cancer. Annu. Rev. Nutr. **18**, 413–440 (1998)
15. Coley, N., Andrieu, S., Gardette. V., Gillette-Guyonnet, S., Sanz, C., Vellas, B., Grand, A.: Dementia prevention: Methodological explanations for inconsistent results. Epidemiol. Rev. 35–66 (2008)
16. Coulibaly, B., Li, G.: Do homeowners increase consumption after the last mortgage payment? An alternative test of the permanent income hypothesis. Rev. Econ. Stat. **88**, 10–19 (2006)
17. Darwin, C.: Letter to Henry Fawcett, 18 September 1861 (1861)
18. Dretske, F.I.: Reasons and falsification. Philos. Q. **15**, 20–34 (1965)
19. Fermi, L., Bernardini, G.: Galileo and the Scientific Revolution. Basic Books, New York (1961). Reprinted Dover, New York (2003)
20. Feyerabend, P.: How to be a good empiricist — a plea for tolerance in matters epistemological. In: Nidditch, P.H. (ed.) The Philosophy of Science. Oxford University Press, New York (1968)
21. Feyerabend, P.: Against Method. Verso, London (1975)
22. Fodor, J.A.: The dogma that didn't bark. Mind **100**, 201–220 (1991)

23. Franklin, A., Howson, C.: Why do scientists prefer to vary their experiments? Stud. Hist. Philos. Sci. **15**, 51–62 (1984)
24. Friedman, M.: A Theory of the Consumption Function. Princeton University Press, Princeton, NJ (1957)
25. Friedman, M.: The Essence of Friedman. Hoover Institution Press, Stanford, CA (1987)
26. Galileo, G.: Two New Sciences (1638). University of Wisconsin Press, Madison (1974)
27. Glazer, A.: Advertising, information and prices. Econ. Inquiry **19**, 661–671 (1981)
28. Gross, D.B., Souleles, N.S.: Do liquidity constraints and interest rates matter for consumer behavior? Evidence from credit card data. Q. J. Econ. **117**, 149–185 (2002)
29. Gruber, J.: Covering the uninsured in the United States. J. Econ. Lit. **46**, 571–606 (2008)
30. Harré, R.: Great Scientific Experiments. Phaidon Press, London (1981). Reprinted Dover, New York (2002)
31. Ho, D.E., Imai, K.: Estimating causal effects of ballot order from a randomized natural experiment: The California alphabet lottery, 1978–2002. Public Opin. Q. **72**, 216–240 (2008)
32. Hubbard, R.L., Craddock, S.G., Flynn, P.M., Anderson, J., Etheridge, R.M.: Overview of 1-year follow-up outcomes in the Drug Abuse Treatment Outcome Study (DATOS). Psychol. Addict. Behav. **11**, 261–278 (1997)
33. Kanny, D., Schieber, R.A., Pryor, V., Kresnow, M.: Effectiveness of a state law mandating use of bicycle helmets among children. Am. J. Epidemiol. **154**, 1072–1076 (2001)
34. Kuhn, E.M., Nie, C.L., O'Brien, M.E., Withers, R.L., Wintemute, G.J., Hargarten, S.W.: Missing the target: A comparison of buyback and fatality related guns. Inj. Prev. **8**, 143–146 (2002)
35. Kuhn, T.: A function for thought experiments. In: Kuhn, T. (ed.) The Essential Tension: Selected Studies in Scientific Tradition and Change, pp. 240–265. University of Chicago Press, Chicago (1979)
36. Larzelere, R.E., Cox, R.B., Swindle, T.M.: Many replications do not causal inferences make: the need for critical replications to test competing explanations of nonrandomized studies. Perspect. Psychol. Sci. **10**, 380–389 (2015)
37. Lumey, L.H., Stein, A.D., Kahn, H.S., van der Pal-de Bruin, K.M., Blauw, G.J., Zybert, P.A., Susser, E.S.: Cohort profile: The Dutch hunger winter families study. Int. J. Epidemiol. **36**, 1196–1204 (2007)
38. Manski, C.F., Pepper, J.V., Thomas, Y.F. (eds): Assessment of Two Cost-Effectiveness Studies on Cocaine Control Policy. National Academies Press, Washington, DC (1999)
39. McCullough, M.L., Giovannucci, E.L.: Diet and cancer prevention. Oncogene **23**, 6349–6364 (2004)
40. Mill, J.S.: A System of Logic: The Principles of Evidence and the Methods of Scientific Investigation. Liberty Fund, Indianapolis (1867)
41. Milyo, J., Waldfogel, J.: The effect of price advertising on prices: Evidence in the wake of 44 Liquormart. Am. Econ. Rev. **89**, 1081–1096 (1999)
42. Mitrou, P.N., Albanes, D., Weinstein, S.J., Pietinen, P., Taylor, P.R., Virtamo, J., Leitzmann, M.F.: A prospective study of dietary calcium, dairy products and prostate cancer risk. Int. J. Cancer **120**, 2466–2473 (2007)
43. Munafo, M.R., Davey Smith, G.: Repeating experiments is not enough. Nature **553**, 399–401 (2018)
44. Neumark, D., Wascher, W.: Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania: Comment. Am. Econ. Rev. **90**, 1362–1396 (2000)
45. Nozick, R.: Philosophical Explanations. Harvard University Press, Cambridge, MA (1981)
46. Park, S.Y., Murphy, S.P., Wilkens, L.R., Stram, D.O., Henderson, B.E., Kolonel, L.N.: Calcium, vitamin D, and dairy product intake and prostate cancer risk — The multiethnic cohort study. Am. J. Epidemiol. **166**, 1259–1269 (2007)
47. Peirce, C.S.: The architecture of theories. In: Charles Sanders Peirce: Selected Writings. Dover, New York (1958)
48. Platt, J.: Strong inference. Science **146**, 347–353 (1964)

49. Plott, C.R., Smith, V.L. (eds): Handbook of Experimental Economics Results. North Holland, New York (2008)
50. Polya, G.: Heuristic reasoning and the theory of probability. Am. Math. Month. **48**, 450–465 (1941)
51. Popper, K.R.: The Logic of Scientific Discovery. Harper and Row, New York (1935 in German, 1959 in English)
52. Quine, W.V.O.: What price bivalence? J. Philos. **78**, 90–95 (1981)
53. Rachlin, H.: Self-control: Beyond commitment (with Discussion). Behav. Brain Sci. **18**, 109–159 (1995)
54. Romer, D.: Advanced Macroeconomics. McGraw-Hill/Irwin, New York (2005)
55. Rosenbaum, P.R.: Choice as an alternative to control in observational studies (with Discussion). Stat. Sci. **14**, 259–304 (1999)
56. Rosenbaum, P.R.: Replicating effects and biases. Am. Stat. **55**, 223–227 (2001)
57. Rosenbaum, P.R.: Reasons for effects. Chance 5–10 (2005)
58. Samuelson, P.A.: Schumpeter as economic theorist. In: Samuelson, P.A. (ed.) Collected Papers, Vol. 5, pp. 201–327. MIT Press, Cambridge, MA (1982, 1986)
59. Schatzkin, A., Gail, M.: The promise and peril of surrogate end points in cancer research. Nat. Rev. Cancer **2**, 19–27 (2002)
60. Sparen, P., Vagero, D., Shestov, D.B., Plavinskaja, S., Parfenova, N., Hoptiar, V., Paturot, D., Galanti, M.R.: Long term mortality after severe starvation during the siege of Leningrad: prospective cohort study. Br. Med. J. **328**, 11–14 (2004)
61. Souleles, N.S.: The response of household consumption to income tax refunds. Am. Econ. Rev. **89**, 947–958 (1999)
62. Staley, K.W.: Robust evidence and secure evidence claims. Philos. Sci. **71**, 467–488. (2004)
63. Surgeon General's Advisory Committee: Smoking and Health. Van Nostrand, Princeton, NJ (1964)
64. Stein, Z., Susser, M., Saenger, G., Marolla, F.: Nutrition and mental performance. Science **178**, 708–713 (1972)
65. Stigler, G.J.: The economics of minimum wage legislation. Am. Econ. Rev. **36**, 358–365 (1946)
66. Susser, E., Hoek, H.W., Brown, A.: Neurodevelopmental disorders after prenatal famine: the story of the Dutch famine study. Am. J. Epidemiol. **147**, 213–216 (1998)
67. Susser, M.: Causal Thinking in the Health Sciences: Concepts and Strategies in Epidemiology. Oxford University Press, New York (1973)
68. Susser, M.: Falsification, verification and causal inference in epidemiology: Reconsideration in the light of Sir Karl Popper's philosophy. In: Susser, M. (ed.) Epidemiology, Health and Society: Selected Papers, pp. 82–93. Oxford University Press, New York (1987)
69. Urmson, J.O.: Aristotle's Ethics. Blackwell, Oxford (1988)
70. Weed, D.L., Hursting, S.D.: Biologic plausibility in causal inference: current method and practice. Am. J. Epidemiol. **147**, 415–425 (1998)
71. Weiss, N.S.: Can the "specificity" of an association be rehabilitated as a basis for supporting a causal hypothesis? Epidemiology **13**, 6–8 (2002)
72. Wellford, C.F., Pepper, J.V., Petrie, C.V., and the Committee to Improve Research and Data on Firearms: Firearms and Violence. National Academies Press, Washington, DC (2004)
73. Williams, B.: Knowledge and reasons. In: Williams, B. (ed.) Philosophy as a Humanistic Discipline, pp. 47–56. Princeton University Press, Princeton, NJ (2006)
74. Wintemute, G.J., Wright, M.A., Drake, C.M., Beaumont, J.J.: Subsequent criminal activity among violent misdemeanants who seek to purchase handguns: Risk factors and effectiveness of denying handgun purchase. J. Am. Med. Assoc. **285**, 1019–1026 (2001)
75. Wolpin, K.I.: The Limits of Inference without Theory. MIT Press, Cambridge, MA (2013)
76. Zellner, A. Readings in Economic Statistics and Econometrics. Little, Brown, Boston (1968)

Chapter 5
Opportunities, Devices, and Instruments

Abstract What features of the design of an observational study affect its ability to distinguish a treatment effect from bias due to an unmeasured covariate u_{ij}? This topic, which is the focus of Part III of the book, is sketched in informal terms in the current chapter. An opportunity is an unusual setting in which there is less confounding with unobserved covariates than occurs in common settings. One opportunity may be the base on which one or more natural experiments are built. A device is information collected in an effort to disambiguate an association that might otherwise be thought to reflect either an effect or a bias. Typical devices include: multiple control groups, outcomes thought to be unaffected by the treatment, coherence among several outcomes, and varied doses of treatment. An instrument is a relatively haphazard nudge towards acceptance of treatment where the nudge itself can affect the outcome only if it prompts acceptance of the treatment. Although competing theories structure design, opportunities, devices, and instruments are ingredients from which designs are built.

5.1 Opportunities

5.1.1 The Well-Ordered World

. . . a nostalgia for caprice and chaos . . .

E.M. Cioran [35, p. 2]

In a well-ordered world, a rationally ordered world, each person would have no choice but to receive the best treatment for that person, the decision being based on undisputed knowledge and expert deliberation. That well-ordered world would be the antithesis of a randomized experiment, in which people receive treatments for no reason at all, simply the flip of a fair coin, the experimental design reflecting our acknowledged ignorance about the best treatment and our determined efforts to reduce that ignorance. In that well-ordered world, it would be difficult to learn anything new. In that world, policy has settled into the "deep slumber of decided

© Springer Nature Switzerland AG 2020

P. R. Rosenbaum, *Design of Observational Studies*, Springer Series in Statistics,
https://doi.org/10.1007/978-3-030-46405-9_5

opinion" [103, p. 42]. Happily for the investigator planning an observational study, we do not live in that well-ordered world.

An observational study may begin with an opportunity, an arbitrary, capricious, chaotic disruption of the ordered world of everyday. The investigator may ask: How can this disruption be put to use? Is there a pair of scientific theories that on an average day do not yield conflicting predictions, but today, just because of this disruption, do yield conflicting predictions? In Chap. 2, randomness was found to be useful. Following Richard Sennett [151], we may ask: Is disorder useful?

5.1.2 Questions

I. Does a low level of ambient light cause auto accidents? It would not do to compare shoppers driving at 11:00 am, commuters driving at dusk, party-goers driving home at 3:00 am; their rates of accidents might differ for reasons other than the difference in ambient light. It would not do to compare the rates of accidents at 5:00 pm in Helsinki and Tel Aviv, because a difference in ambient light is one difference among many others. What would you compare?

II. If murderers, rapists, and armed robbers condemned to death were not executed, would they do greater injury to other inmates and prison staff than is done by others incarcerated for violent crimes? What would you compare?

III. At present in the USA, married couples pay taxes jointly on the sum of their incomes, where the total income matters, but the division of income between husband and wife does not. Also, the tax rates are progressive, with higher tax rates at higher incomes. Under such a system, if a husband and wife have very different incomes, the tax disincentives to work may be quite different than if they were not married. It is natural to ask: What effect does joint taxation have on the labor supply of husbands and wives? How would you study the effects of joint taxation on labor supply given that the whole country is subject to joint taxation?

IV. Does growing up in a poor neighborhood depress later adult earnings? The question asks for a separation of the effects of the neighborhood from any attributes of the individual's own family circumstances. What would you compare?

5.1.3 Solutions

I. You are looking for a change in ambient light with no other consequential change. When does that happen? It happens twice a year with the switch into and out of daylight savings time. John Sullivan and Michael Flannagan [160] looked at change in fatal auto accident rates at dawn and dusk in the weeks before and after the discontinuity in ambient light produced by the switch into and out of

daylight savings time. Does increased or reduced sleep affect auto accident rates? In a different use of the change into or out of daylight savings time, Mats Lambe and Peter Cummings [87] looked at Mondays before and after the switch, noting that on the Monday immediately after the switch, people often have added or subtracted an hour of sleep.

The use of daylight savings time to vary either ambient light or sleep duration is an example of a discontinuity design, in which a treatment shifts abruptly and discontinuously on some dimension, here time, while the sources of bias that are of greatest concern are likely to shift gradually and continuously, so the bias in estimating the treatment effect is small near the discontinuity. Donald Thistlethwaite and Donald Campbell [164] were the first to see discontinuities as opportunities for observational studies; see also [22, 23, 26, 38, 59, 70].

II. Generally, it is not easy to study the behavior that a person would have exhibited had they not been executed. In the case of Furman v. Georgia in 1972, the U.S. Supreme Court found that the then current methods of implementing the death penalty were "cruel and unusual" and therefore in violation of the U.S. Constitution, invalidating the death sentences of more than 600 prisoners facing execution. Marquart and Sorensen [98] looked at 47 Texas inmates whose sentences were commuted to life imprisonment, examining their behavior from 1973 to 1986, compared with a cohort of similar violent offenders not on death row.

Supreme Court decisions are not infrequently seen as opportunities; e.g., [104]. For present purposes, it makes little difference whether we view the Supreme Court as capricious in striking down legislation or legislators as capricious in ignoring the strictures of the Constitution; not wisdom, but sharp abrupt change is sought.

III. Today, the U.S. tax code is well-ordered in the limited sense that the whole country is subject to joint taxation, but it was not always so. Prior to the Revenue Act of 1948, married couples paid joint taxes on their combined incomes in states with community property laws and separate taxes on their individual incomes in states without such laws. Sara LaLumia [86] compared the change in the labor supplied by husbands and wives, from before the Revenue Act of 1948 to after, in states with and without community property laws.

IV. The public housing program in Toronto assigned families to locations based on vacancies at the time a family reached the top of the waiting list, and Philip Oreopoulos [108] used this relatively haphazard arrangement to study neighborhood effects on adult earnings. He wrote:

> This paper is the first to examine the effects of the neighborhood on the long-run labor market outcomes of adults who were assigned as children to different residential housing projects in Toronto ... All families in the Toronto program were assigned to various housing projects throughout the city at the time they reach the top of the waiting list ... [F]amilies cannot specify a preference ... The Toronto housing program also permits comparison across a wide variety of subsidized housing projects ... [some located] in central downtown, while others are in middle-income areas in the suburbs.

The Toronto housing program is an opportunity to the extent that position on a waiting list predicts earnings later in life only because it affects the assigned neighborhood. The Toronto housing program is an opportunity to the extent that something consequential (your neighborhood) was determined by something inconsequential (your position on a waiting list).

5.2 Devices

5.2.1 Disambiguation

In his President's Address to the Royal Society of Medicine, Sir Austin Bradford Hill [63] asked:

> Our observations reveal an association between two variables, perfectly clear-cut and beyond what we would care to attribute to the play of chance. What aspects of that association should we especially consider before deciding that the most likely interpretation of it is causation?

Association does not imply causation: an association between treatment and outcome is ambiguous, possibly an effect caused by the treatment, possibly a bias from comparing people who are not comparable despite looking comparable. "Disambiguate" is a word with an ugly sound but the right attitude. Ambiguity is opposed, perhaps ultimately defeated, by activity specifically targeted at resolving ambiguity.[1] Devices enlarge the set of considered associations with a view to disambiguating the association between treatment and outcome. Donald Campbell's paper [30] of 1957 was one of the first to systematically consider the role of devices.

5.2.2 Multiple Control Groups

In the absence of random assignment, the mere fact that controls escaped treatment is prima facie evidence that they are not comparable to treated subjects. How did they escape treatment? Perhaps the controls were denied treatment, or declined the offer of treatment, or lived too far away to receive treatment, or were studied before treatment became available. Depending upon the context, these reasons for escaping treatment may appear consequential or innocuous. How might one obtain relevant evidence to supplement mere appearances?

If there are several ways to escape treatment, then several control groups are possible. Direct evidence about bias from unmeasured covariates is sometimes available, depending upon the context, by using more than one control group in

[1] Citing Bentham, the Oxford English Dictionary writes: "'disambiguate': verb, to remove ambiguity from."

Table 5.1 Comparison of micronuclei frequency among lead-zinc miners and two control groups

Group	Label	n	Mean	SE	t vs LP	t vs SR
Local population	LP	57	6.005	0.377	–	−0.98
Slovene residents	SR	61	6.400	0.143	0.98	–
Mine workers	MW	67	14.456	0.479	13.87	16.13

The two-sample unpooled t-statistics are given together with the standard errors (SE) of the means

an observational study. For instance, one control group might consist of people who declined treatment and another control group of people who were denied treatment.

For example, in 2005, Marjan Bilban and Cvetka Jakopin [25] asked whether lead and zinc miners exposed to radon gas and heavy metals were suffering from excess levels of genetic damage. They compared 70 Slovene lead-zinc miners at one mine with two control groups, one consisting of local residents close to the mine, the other consisting of other Slovene residents who lived at a considerable distance from the mine. The mine itself was a source of lead pollution in the local environment. Bilban and Jakopin looked at several standard measures in genetic toxicology, including the frequency of micronuclei. In this assay [45], a blood sample is drawn and blood lymphocytes are cultured. In a normal cell division, the genetic material in one cell divides to become the genetic material in the two nuclei of two new separate cells. If this process does not go well, each new cell may not have a single nucleus, but may have bits of genetic material scattered in several micronuclei. The frequency of micronuclei (MN) is the measure of genetic damage; specifically, Bilban and Jakopin looked at 500 binuclear cells per person and recorded the total number of micronuclei in the 500 cells.

Table 5.1 displays Bilban and Jakopin's [25, Table IV] data for micronuclei among miners and the two control groups. It is clear that the miners have higher frequencies of micronuclei than both control groups, which do not differ greatly from each other.[2] Another example using two control groups is discussed in Sect. 12.3.

The advantages and limitations of multiple control groups may be developed in formal statistical terms, but several issues are fairly clear without formalities. First, if two control groups have differing outcomes after adjustment for observed covariates, then that cannot be due to an effect caused by the treatment, and it

[2]The analysis as I have done it in Table 5.1 is unproblematic because the situation is so dramatic, with t's either above 10 or below 1. In less dramatic situations, this analysis is not appropriate, for several reasons. First, a more powerful test for effect would use both control groups at once. Second, Table 5.1 performs several tests with no effort to control the rate of misstatements by these tests. Third, Table 5.1 takes the absence of a difference between the two control groups as supporting their comparability, but the failure to reject a null hypothesis is not evidence in favor of that hypothesis. For instance, the two control groups might not differ significantly because of limited power, or else they might differ significantly but the differences might be too small to invalidate their usefulness in the main comparison with the treated group. These issues will be discussed with greater care in Sect. 23.3.

must indicate that at least one and perhaps both control groups are not suitable as control groups, perhaps that they differ with respect to unmeasured covariates. If the two control groups differ from each other with respect to relevant unmeasured covariates, then at least one of the two control groups must differ from the treated group with respect to these covariates.

Second, two control groups can be useful only if they differ in some useful way. If having two control groups were of value merely because there are two instead of one, then any one control group could be divided in half at random to produce two groups, but of course that would provide no new information about unmeasured covariates. The two control groups in Table 5.1 help to indicate that there is nothing special about the region immediately around the mine; only miners, not local residents, have elevated MN levels. Donald Campbell [31] argued that it is wise to select two control groups so they differ with respect to a covariate that, though unmeasured, is known to differ substantially in the two groups. Similar outcomes in two such control groups provide evidence that imbalances in the unmeasured covariate are not responsible for treatment-vs-control differences in outcomes; see [117, 119] for discussion of this in terms of the power and unbiasedness of tests for bias from unmeasured covariates. Following M. Bitterman, Campbell refers to this as "control by systematic variation:" the unmeasured covariate is systematically varied without producing substantial variation in the outcome.

Third, similar outcomes in two control groups do not ensure that treatment-versus-control comparisons are unbiased. The two control groups may both be biased in the same way. For instance, if a willingness to work in a lead-zinc mine is associated with a general lack of concern about health hazards, one might discover that miners are more prone to tolerate other health hazards, say cigarette smoke and excessive alcohol, than are other Slovenes living near or far from the mine. A comparison of two control groups may have considerable power to detect certain specific biases—biases from covariates that have been systematically varied—and virtually no power to detect other biases—biases that are present in a similar way in both control groups [117, 119]. This is a very simple case of a general issue discussed carefully by Dylan Small [154].

5.2.3 Coherence Among Several Outcomes

Austin Bradford Hill [63] in the quotation in Sect. 5.2.1 argued that there were aspects of the observed association that might aid in distinguishing an actual treatment effect from mere failure to adjust for some covariate that was not measured. Section 5.2.2 considered multiple control groups. In Sect. 3.4, it was seen that studies vary in their sensitivity to bias from unmeasured covariates: small biases, measured by Γ, could explain the observed association between treatment and outcome in some studies, but only large biases do so in other studies. So the degree of sensitivity to unmeasured biases is another aspect of the observed

Table 5.2 Four of the 86 pairs of two Israeli schools in Angrist and Lavy's study of test performance and class size manipulated using Maimonides' rule

Pair i	School j	Percentage Disadvantaged x	Cohort Size	Number of Classes	Class Size	Average Math Score	Average Verbal Score
1	1	9	46	2	23.0	72.1	81.1
1	2	8	40	1	40.0	63.1	79.4
2	1	1	45	2	22.5	78.5	85.5
2	2	1	32	1	32.0	68.1	75.7
3	1	0	47	2	23.5	78.1	80.0
3	2	0	40	2	20.0	64.4	80.4
4	1	1	45	2	22.5	76.4	82.9
4	2	1	33	1	33.0	67.0	84.0

Pair $j = 3$ violated Maimonides' rule, dividing the cohort for school $j = 2$ which had a cohort size of 40

association that is relevant to distinguishing treatment effects and biases. Are there additional aspects to consider?

Table 5.2 displays data for the fifth grade for four of the 86 pairs of Israeli schools in Fig. 1.1 from Angrist and Lavy's [4] study of academic test performance and class size manipulated by Maimonides' rule. Recall from Sect. 1.3 that schools were paired for the percentage of disadvantaged students in the school, x, and the outcomes were the average math and verbal test scores for the fifth grade. The middle portion of Table 5.2 shows Maimonides' rule in imperfect operation. School $(i, j) = (1, 1)$ had 46 students in the fifth grade and was divided to form two classes with an average size of 23, while school $(i, j) = (1, 2)$ had 40 students in the fifth grade and was not divided, yielding one class with 40 students. Pair $i = 3$ violates Maimonides' rule because the cohort of size 40 in school $(i, j) = (3, 2)$ was divided.

Table 5.2 suggests several questions. The first question, discussed in the current section, concerns "coherence."[3] As is often the case, the most compelling argument in support of the importance of "coherence" comes from considering "incoherence." Suppose that larger cohorts divided into smaller classes had decidedly superior results in mathematics when compared with smaller cohorts taught in larger classes, but also decidedly inferior results in the verbal test. That result would be incoherent. Faced with an incoherent result of this sort, the investigator would be hard pressed to argue that smaller classes produce superior academic achievement. In fact, Fig. 1.1 suggests gains in both math and verbal test scores, a coherent result. If incoherence presents a substantial obstacle to a claim that the treatment caused its ostensible

[3]Hill [63] used the attractive term "coherence" but did not give it a precise meaning. Campbell [32] used the term "multiple operationalism" in a more technical sense, one that is quite consistent with the discussion in the current section. Trochim [165] uses the term "pattern matching" in a similar way. Reynolds and West [113] present a compelling application.

effects, then the absence of incoherence—that is, coherence—should entail some strengthening of that claim. Can this intuition be formalized?

Recall from Sect. 2.4.1 the hypothesis of an additive treatment effect. If indeed the treatment had additive effects, τ_{math} on the math test, τ_{verb} on the verbal test, then a claim that smaller class sizes improve test performance is the claim that $\tau_{math} \geq 0$ and $\tau_{verb} \geq 0$ with at least one strict inequality. To say that smaller classes cause a decisive improvement in both mathematics and verbal test performance is to say that both τ_{math} and τ_{verb} are substantially positive. A gain in math and a loss in verbal test performance, $\tau_{math} > 0$ and $\tau_{verb} < 0$, is not logically impossible, but it would not be compatible with the anticipated benefit from smaller class sizes. The policy question is whether smaller classes confer benefits commensurate with their unambiguously higher costs, and for that the relevant question is whether both τ_{math} and τ_{verb} are substantially positive. Viewed geometrically, a gain in mathematics is half of the line, $\tau_{math} > 0$, but a gain in both mathematics and verbal scores is a quarter of the plane, $(\tau_{math}, \tau_{verb}) > (0, 0)$, so the latter is a more focused hypothesis.

In the previous paragraph I have made an argument that a relevant treatment effect has a certain specific form. You may or may not have found that argument compelling. No matter. For present purposes, the issue is not whether the argument in the previous paragraph is compelling to you in this particular case, but rather that any claim of coherence or incoherence depends, explicitly or implicitly, on an argument of this sort. A claim is made that an actual treatment effect, or a policy-relevant treatment effect, or a useful treatment effect must have a certain form. Coherence then means a pattern of observed associations compatible with this anticipated form, and incoherence means a pattern of observed associations incompatible with this form. Claims of coherence or incoherence are arguable to the extent that the anticipated form of treatment effect is arguable.

A simple way to look for an increase in both math and verbal scores is to add together the two signed rank statistics for the math and verbal scores[4]; this is one instance of the "coherent signed rank statistic" [124] and [129, §9]. Although discussed in technical terms in Sect. 18.2, the intuition behind the coherent signed rank statistic is clear without technical details. If larger cohorts divided in half tend to have better math and verbal scores than smaller cohorts, both signed rank statistics will tend to be large, and the coherent statistic will be extremely large. If gains in math scores are offset by losses in verbal scores, the sum of the two signed rank statistics will not be especially large.

[4]In an obvious way, in adding the two signed rank statistics, one is committing oneself to a particular direction of effect for the two outcomes. If one anticipated gains in math scores together with declines in verbal scores, one might replace the verbal score by their negation before summing the signed rank statistics. Because the two outcomes are ranked separately, approximately equal weight is being given to each of the outcomes. The coherent signed rank statistic may be used with more than two oriented outcomes. It may also be adjusted to include varied doses of treatment [124].

Table 5.3 Sensitivity analysis and coherence in Angrist and Lavy's study of academic test performance and class size manipulated by Maimonides' rule

Γ	Math	Verbal	Coherent
1.00	0.0012	0.00037	0.00018
1.40	0.043	0.020	0.011
1.45	0.057	0.027	0.015
1.55	0.092	0.047	0.026
1.65	0.138	0.075	0.043
1.70	0.164	0.092	0.054

The table gives the upper bound on the one-sided P-value for the math test score, the verbal test score and their coherent combination. In this instance, a larger bias, say $\Gamma = 1.7$, is needed to explain away the coherent association, whereas the association with the math test score could be explained by $\Gamma = 1.45$

We may conduct a sensitivity analysis for the coherent signed rank statistic. See [124] and [129, §9]. The intuition here is that a coherent result will be judged less sensitive to unobserved bias if the coherent signed rank statistic is used, whereas an incoherent result will be more sensitive to unobserved bias. Section 18.2 will develop some of the formalities that justify this intuition. The computations are very similar in form to those in Sect. 3.5 and equally simple to perform; see [124] and [129, §9].

Table 5.3 shows the sensitivity analysis for the signed rank statistic applied to the math score alone, to the verbal score alone, and to the coherent signed rank statistic that combines them. Table 5.3 presents the upper bound on the one-sided P-value for testing the null hypothesis of no treatment effect against an increase in test performance. The lower bound on the P-value is highly significant in all cases and is not presented. The math score becomes sensitive to unobserved bias at about $\Gamma = 1.45$, the verbal score at a little more than $\Gamma = 1.55$ and the coherent statistic at about $\Gamma = 1.70$. In brief, a noticeably larger bias, Γ, would need to be present to explain away the coherent pattern of associations than would be needed to explain away the ostensible effect on either math or verbal scores taken in isolation. See Sect. 23.4 for further discussion of this analysis.

The results in Table 5.3 for Israeli schools are more sensitive to unobserved biases than the results in Table 3.2 for welders. That is an unambiguous fact clearly visible in the two sets of data. However, the contexts are different, and the contexts are relevant to thinking about what biases are plausible. Angrist and Lavy [4] worked hard to prevent unmeasured biases: the treated and control schools in Fig. 1.1 and Table 5.2 were selected to be similar in terms of percentage disadvantaged, x, and to differ slightly in fifth grade cohort size on opposite sides of Maimonides' cutpoint of 40 students. Could this comparison omit a covariate u strongly predictive of both math and verbal test performance and about 1.7 times more common among schools with a slightly larger fifth grade cohort? Certainly, that remains a logical

possibility, but Angrist and Lavy's results in Fig. 1.1 are not extremely fragile given the magnitude of bias that seems plausible here.

5.2.4 Known Effects

Unaffected Outcomes or Control Outcomes

In Sect. 5.2.3, coherence referred to the possibility that the treatment was thought to affect several outcomes in known directions. It may happen, however, that we think a treatment will affect one outcome and not affect another, and we wish to exploit the anticipated absence of effect to provide information about unmeasured biases [100, 113, 114, 118, 119, 172, 173]. A control group is a group of subjects known to be unaffected by the treatment; by analogy, an outcome known to be unaffected by the treatment is sometimes called a "control outcome" or a "control construct." A control outcome is sometimes linked to Hill's [63] notion of the "specificity" of a treatment effect [118, 172].[5] In particular, McKillip [100] argues that under some circumstances, a control outcome might suffice in the absence of a control group.

An argument involving a control outcome appeared in Sect. 4.3 in the study by Wintemute et al. [174] of the effects in California of the introduction of a law in 1991 denying handguns to individuals convicted of violent misdemeanors. Recall that they compared individuals who had convictions for violent misdemeanors and who applied to purchase a handgun in either 1989–1990, before the law, or 1991, after the law had taken effect. Also, recall that they found lower rates of gun and violent crime when the law was in effect. Is this an effect caused by the law? Over time, there could be shifts in crime rates with, say, shifts in demographics or economic conditions. It is not clear how or why restrictions on handgun purchases would affect nonviolent crimes. Wintemute et al. [174] also looked at rates of nonviolent and nongun crime, finding little change in these rates following the change in law. In their study, where an effect is plausible, treatment and outcome are associated; where an effect is not particularly plausible, treatment and outcome are not associated.

In a similar manner, Sadik Khuder et al. [85] study the impact of a smoking ban in workplaces and in public areas on the rate of hospital admission for coronary heart disease. As a control outcome, they used admissions for non-smoking related diseases.

Although the role of known effects may be developed in formal terms, the main issues are fairly clear without technical detail. To be useful, an unaffected outcome must have certain properties. If an unaffected outcome were useful merely because it is unaffected, then such an outcome could always be created artificially using

[5]Specificity of a treatment effect in Hill's sense [63] is sometimes understood as referring to the number of outcomes associated with the treatment, but more recent work has emphasized the absence of associations with outcomes the treatment is not expected to affect [118, 172].

random numbers; however, that could not provide insight into unmeasured biases. To be useful, an unaffected outcome must be associated with some unmeasured covariate. Plausibly, trends in nonviolent crime or nongun crime are associated with general trends in lawless behavior, and therefore are useful in distinguishing an effect of the law from a general increase or decrease in crime. It is not difficult to show that an unaffected outcome can provide a consistent and unbiased test[6] of imbalance in an unmeasured covariate with which it is associated; see [118, 119] and [129, §6]. A substantial imbalance in an unmeasured covariate may yield only a faint echo in an unaffected outcome if the association between them is weak.[7] An unaffected outcome may guide the scope of a sensitivity analysis [121].

Bias of Known Direction

Claims to know the direction of the bias induced by an unmeasured covariate are common in discussions of the findings of observational studies. To the extent that such claims are warranted, they may disambiguate certain associations between treatment and outcome. Did adherence to the gold standard lengthen the Great Depression? Building upon work by Milton Friedman and Anna Schwartz [50] about the USA, studies by E.U. Choudhri and L.A. Kochin [34] and B. Eichengreen and J. Sachs [41] compared nations that were not on the gold standard, that left the gold standard at various times, or that remained on the gold standard. Discussing these studies, Benjamin Bernanke [24] wrote:

> [B]y 1935 countries that had left gold relatively early had largely recovered from the Depression, while Gold Block countries remained at low levels of output and employment. ... If choices of exchange-rate regime were random, these results would leave little doubt ... Of course, in practice the decision about whether to leave the gold standard was endogenous to a degree ... In fact, these results are very unlikely to be spurious ... [A]ny bias created by endogeneity of the decision to leave gold would appear to go the wrong way, as it were, to explain the facts: The presumption is that economically weaker countries,

[6]Consistency and unbiasedness are two concepts of minimal competence for a test of a null hypothesis H_0 against an alternative hypothesis H_A. Consistency says the test would work if the sample size were large enough. Unbiasedness says the test is oriented in the correct direction in samples of all sizes. One would be hard pressed to say the test is actually a test of H_0 against H_A if consistency and unbiasedness failed in a material way. To be a 5% *level* test of H_0, the chance of a P-value less than 0.05 must be at most 5% when H_0 is true. The *power of a test* of a null hypothesis, H_0, against an alternative hypothesis, H_A, is the probability that H_0 will be rejected when H_A is true. If the test is performed at the 5% level, then the power of the test is the probability of a P-value less than or equal to 0.05 when H_0 is false and H_A is true instead. We would like the power to be high. The test is *consistent* against H_A if the power increases to 1 as the sample size increases—that is, rejection of H_0 in favor of H_A is nearly certain if H_A is true and the sample size is large enough. The test is an *unbiased* test of H_0 against H_A if the power is at least equal to the level whenever H_A is true. If the test is performed at the 5% level, then it is unbiased against H_A if the power is at least 5% when H_A is true.

[7]More precisely, the Kullback–Leibler information in the unaffected outcome is never greater, and is typically much smaller, than the information in the unmeasured covariate itself [118].

> or those suffering the deepest depressions, would be the first to devalue or abandon gold. Yet the evidence is that countries leaving gold recovered substantially more rapidly and vigorously than those who did not. Hence, any correction for endogeneity … should tend to strengthen the association of economic expansion and the abandonment of gold.

Bernanke's claim is typical of claims to know the direction of a particular bias. The claim is not put forward in the language of absolute certainty. Rather the claim is put forward as a clarification of the logical and yet somewhat implausible consequences of the claim that the association between treatment and outcome was produced by a bias of a particular kind. To assert the claimed bias is to assert also its necessary but implausible implications.

A fairly risky but sometimes compelling study design exploits a claim to know that the most plausible bias runs counter to the claimed effects of the treatment. In this design, two groups are compared that are known to be incomparable, but incomparable in a direction that would tend to mask an actual effect rather than create a spurious one. The logic behind this design is valid: if the bias runs counter to the anticipated effect, and the bias is ignored, inferences about the effect will be conservative, so the bias will not lead to spurious rejection of no effect in favor of the anticipated effect; see [118, 119] and [129, §6]. Nonetheless, the design is risky because a bias in one direction combined with an effect in the other may cancel, so an actual effect is missed. Moreover, the claim to know the direction of the bias may be disputed.

Does disability insurance discourage work? In one study exploiting unmeasured biases of known direction, John Bound [27] compared successful and unsuccessful applicants for U.S. Social Security disability insurance. His premise was that successful applicants were often more severely disabled than unsuccessful applicants, that is, that the U.S. Social Security Administration was employing some reasonable criteria in sorting the applications into two piles. In this case, the incentive to not work was given to individuals who were, presumably, less able to work. Presumably, the difference in work behavior between successful and unsuccessful applicants overstates the effect of the incentive, because it combines the impact of the incentive and greater disability. In fact, Bound [27] found that relatively few of the unsuccessful applicants returned to work, and on that basis claimed that the incentive effects were not extremely large.

Do laws that prevent handgun purchases by convicted felons prevent violent crimes? What would be a suitable control group for comparison with convicted felons? Mona Wright et al. [176] conducted a study exploiting a bias of known direction. They compared convicted felons whose attempt to purchase a handgun was denied with persons arrested for felonies but not convicted whose purchase was permitted. Presumably, some of those arrested but not convicted were innocent, so presumably the group of purchasers in their study contains fewer individuals who have previously committed felony offenses. Though the most plausible bias works in the opposite direction, Wright et al. found that, adjusting for age, purchasers were 13% more likely than denied felons to be arrested for gun crimes in the following 3 years [176, relative risk 1.13 in Table 1].

5.2.5 *Doses of Treatment*

Does a Dose–Response Relationship Strengthen Causal Inference?

Much has been written about doses of treatment and their relationship to claims of cause and effect. Many such claims contradict other related claims, or appear to. Section 18.4 will attempt to sort out some of these claims in formal terms. Here, some of the claims are mentioned and an example (guided by Sect. 18.4) is presented. (A different aspect of doses is discussed in Sect. 12.3.)

The most familiar claim is due to Austin Bradford Hill in the paper quoted previously. Hill [63, p. 298] wrote:

> [I]f the association is one which can reveal a biological gradient, or dose–response curve, then we should look most carefully for such evidence. For instance, the fact that the death rate from cancer of the lung rises linearly with the number of cigarettes smoked daily, adds a very great deal to the simpler evidence that cigarette smokers have a higher death rate than non-smokers.

Hill's claim is often disputed. Kenneth Rothman [145, p. 18] wrote:

> Some causal associations, however, show no apparent trend of effect with dose; an example is the association between DES and adenocarcinoma of the vagina . . . Associations that do show a dose–response trend are not necessarily causal; confounding can result in such a trend between a noncausal risk factor and disease if the confounding factor itself demonstrates a biologic gradient in its relation with disease.

In a subtle reinterpretation of Hill's notion of dose and response, Noel Weiss [171, p. 488] says much the same:

> [O]ne or more confounding factors can be related closely enough to both exposure and disease to give rise to [a dose response relationship] in the absence of cause and effect.

These remarks speak of a dose–response relationship as something that is present or absent, with contributions to the strength of evidence that are correspondingly present or absent. Is this the best way to speak about either a dose–response relationship or the strength of evidence? We might ask: would an analysis that took account of doses of treatment exhibit less sensitivity to unmeasured biases? The answer is accessible by analytical means in any one study [131, 137], and the answer varies from one study to the next. Hill makes a positive reference to smoking and cancer of the lung, and Rothman makes a negative reference to DES and cancer of the vagina, but the existence of both positive and negative examples is consistent with the notion that doses sometimes reduce sensitivity to unobserved biases.

Cochran [36] argued that observational studies should be patterned after simple experiments, and the good, standard advice about clinical trials is to compare two treatments that are as different as possible; see Peto, Pike, Armitage, Breslow, Cox, Howard, Mantel, McPherson, Peto, and Smith [110, p. 590]. Of course, this entails widely separated doses with no attempt to discover a graduated continuum of dose and response. Is a graduated continuum the relevant issue? Or is it simply important to anticipate negligible effects among treated subjects receiving negligible doses? The issue turns out to be quite unambiguous; see Sect. 18.4 and [132].

Doses may be suspect in one sense or another. In Angrist and Lavy's [4] study of class size and educational test performance, cohort size was thought to be haphazard, but class size—the dose of treatment—was thought to be haphazard only to the extent that it was determined by cohort size using Maimonides' rule. The distinction between treated and control may be free of error, while the dose magnitudes may be subject to errors of measurement [134, 156]. In both instances, suspect doses may require instruments if bias is to be avoided; see Sect. 5.3. Sometimes we mistakenly think of a dose as an aspect of the treatment inflicted upon an individual, when in fact the doses individuals receive are partly or entirely reactions to the needs or tolerances of those individuals [138].

It is time to consider an example.

Genetic Damage from Paint and Paint Thinners

Professional painters are exposed to a variety of potential hazards in paint and paint thinners, including organic solvents and lead. Do such exposures cause genetic damage? Pinto et al. [112] compared male professional painters in Yucatan, Mexico, with male clerks, matched for age, with respect to the frequency of micronuclei in 3000 oral epithelial cells scraped from the cheek. See Sect. 5.2.2 for a discussion of the micronucleus assay. The painters were public building painters who worked without masks or gloves, and the dose d_i of their exposure[8] is the number of years of employment as a painter, recorded on a logarithmic scale, specifically $\log_2$ (years), so $\log_2\left(2^k\right) = k$, and in particular, $\log_2(4) = 2$ and $\log_2(32) = 5$. Pinto et al.'s [112] data appear in Table 5.4 and Fig. 5.1. Obviously, age and years as a painter are highly correlated, but the controls are closely matched for age. In Fig. 5.1, painters have higher frequencies of micronuclei, and the difference between painters and matched controls increases with $\log_2$ (years).

Recall from Sect. 2.3.3 that Wilcoxon's signed rank statistic is $T = \sum_{i=1}^{I} \operatorname{sgn}(Y_i) \cdot q_i$ where $\operatorname{sgn}(a) = 1$ if $a > 0$, $\operatorname{sgn}(a) = 0$ if $a \leq 0$, and q_i is the rank of $|Y_i|$. The dose-weighted signed rank statistic [124, 131, 166] with doses d_i, $i = 1, \ldots, I$, is similar except it weights the pairs i by the doses d_i, $T_{\text{dose}} = \sum_{i=1}^{I} \operatorname{sgn}(Y_i) \cdot q_i \cdot d_i$. An alternative statistic is the signed rank statistic applied to the subset of pairs with large doses, say $d_i \geq \widetilde{d}$, that is,

$$T_{\text{high}} = \sum_{\{i:\, d_i \geq \widetilde{d}\}} \operatorname{sgn}(Y_i) \cdot q_i. \tag{5.1}$$

[8]When $\mathscr{F}$ was introduced in Chap. 2, treatment was applied at a single dose, and so doses were not mentioned. In general, if there are fixed doses, one dose d_i for each pair i, then the doses are also part of $\mathscr{F}$. Because previous discussions involving $\mathscr{F}$ had a single dose, we may adopt the new definition that includes doses in $\mathscr{F}$ without altering the content of those previous discussions.

Table 5.4 Micronnuclei per 1000 cells in 22 male painters and 22 male clerks matched for age

Pair	Painter Age	Control Age	Years Painting	$\log_2$(years)	Painter Micronuclei	Control Micronuclei
1	18	18	1.6	0.68	0.32	0.00
2	20	20	1.6	0.68	0.00	0.00
3	40	39	1.6	0.68	0.00	0.00
4	29	29	1.8	0.85	0.32	0.32
5	20	18	2.0	1.00	0.00	0.65
6	26	27	2.0	1.00	0.00	0.00
7	23	23	3.0	1.58	0.95	0.00
8	30	30	3.0	1.58	0.00	0.62
9	31	31	3.5	1.81	0.33	1.96
10	52	51	3.6	1.85	1.99	0.00
11	22	22	4.0	2.00	0.99	0.33
12	47	47	4.0	2.00	1.54	0.00
13	40	39	4.6	2.20	0.33	0.00
14	22	22	4.7	2.23	0.66	0.00
15	23	23	5.0	2.32	0.00	0.00
16	42	42	5.0	2.32	1.63	0.66
17	35	36	8.0	3.00	0.65	0.00
18	48	49	8.0	3.00	1.64	0.64
19	60	58	8.6	3.10	4.84	0.00
20	62	64	11.0	3.46	5.35	1.30
21	41	40	25.0	4.64	1.99	1.33
22	60	63	40.0	5.32	2.89	0.32

For painters, the number of years worked as a painter is recorded, along with its logarithm to the base of 2. Pairs are sorted by increasing years. A line separates painters with less than 4 years of work as a painter

In the example, $\widetilde{d}$ is set at the median dose, here, 4 years or $\widetilde{d} = \log_2(4) = 2$; see the vertical line in Fig. 5.1. The sensitivity analysis for these statistics is easy to perform and closely parallels the sensitivity analysis for the signed rank statistic in Sect. 3.5; see [124, 131].

Table 5.5 asks whether the dose–response pattern in Fig. 5.1 reduced sensitivity to bias from failure to adjust for an unmeasured covariate u_{ij}. In this instance, weighting by doses, T_{dose}, is marginally less sensitive to unmeasured bias than ignoring doses, T, but using only the 12 pairs with high doses, T_{high}, is substantially less sensitive to unmeasured bias. Specifically, a bias of $\Gamma = 2$ could just explain the association between professional painting and micronuclei frequency if that association is measured ignoring doses using Wilcoxon's signed rank statistic, T, because the upper bound on the one-sided P-value is 0.064. A bias of $\Gamma = 2.5$ could just explain the association between treatment and response if the dose-weighted signed rank statistic, T_{dose}, is used. In contrast, even a bias of $\Gamma = 3.3$ could not

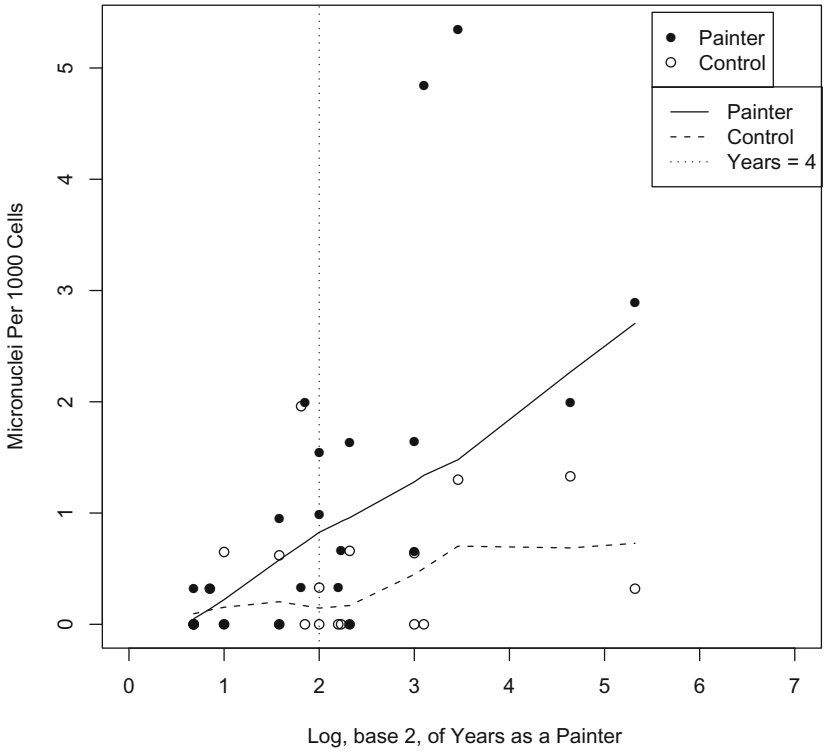

Fig. 5.1 Micronuclei in 22 male painters and 22 male controls matched for age plotted against $\log_2(\text{years})$ where years is the number of years of employment as a painter for the painter in each matched pair. The two curves are lowess smooths. The vertical line is at the median of 4 years or $\log_2(\text{years}) = 2$

Table 5.5 Did a dose–response relationship reduce sensitivity to unmeasured covariates?

Γ	Ignoring doses, T	Weighting by doses, T_{dose}	Using only the 12 high-dose pairs, T_{high}
1	0.0032	0.0025	0.0012
2	0.064	0.038	0.016
2.5	0.12	0.067	0.028
3.3	0.22	0.12	0.048

Tests of the null hypothesis of no treatment effect on micronuclei frequency in 22 male painters and 22 male clerks matched for age. Comparison of the upper bounds on the one-sided P-value using Wilcoxon's signed rank statistic, the dose-weighted signed rank statistic, and the signed rank statistic applied to high-dose pairs. The dose, d_i, is $\log_2(\text{years})$. The signed rank statistic applied to the high-dose pairs used 12 of the 22 pairs in which the painter had worked for at least 4 years as a painter, $d_i = \log_2(\text{years}) \geq 2$, discarding ten pairs

explain the high levels of micronuclei in the 12 pairs with at least 4 years of work as a painter, because the upper bound on the one-sided significance level is 0.048.

A dose–response relationship reduces sensitivity to bias from unmeasured covariates in some studies but not in others [131]. A mere upward trend in a plot or table does not decide the issue, whereas Table 5.5 is decisive. The relative performance of T, T_{dose}, and T_{high} are examined further in Sect. 18.4 and [136]; see also [132].

5.2.6 Differential Effects and Generic Biases

What Are Differential Effects?

If there are two treatments, say A and B, each of which may be applied or withheld, then there is a 2×2 factorial arrangement of treatments in four cells. Extending the notation of Chap. 3, person ℓ falls in one of four cells: (1) neither A nor B ($Z_{A\ell} = 0$, $Z_{B\ell} = 0$), (2) A but not B, ($Z_{A\ell} = 1$, $Z_{B\ell} = 0$), (3) B but not A, ($Z_{A\ell} = 0$, $Z_{B\ell} = 1$), and (4) both A and B, ($Z_{A\ell} = 1$, $Z_{B\ell} = 1$). If subjects ℓ were randomized to the cells of this 2×2 design, with $Z_{A\ell}$ and $Z_{B\ell}$ independently determined by the flips of a fair coin, then there are several comparisons or contrasts that may be examined, and because randomization would prevent bias in all such comparisons, the choice of comparison would reflect only our interests or considerations of parsimony. When randomization is not used, the situation is different. Depending upon the context, it may happen that certain comparisons are less susceptible than others to certain types of biases.

The differential effect [135] of the two treatments is the effect of giving one treatment in lieu of the other. It is a comparison of two of the cells of the 2×2 factorial design, but not one of the most familiar comparisons. Specifically, the differential effect compares cells (2) and (3) above, that is, A but not B, ($Z_{A\ell} = 1$, $Z_{B\ell} = 0$) versus B but not A, ($Z_{A\ell} = 0$, $Z_{B\ell} = 1$). In a randomized experiment, there is no more interest in this comparison than in other comparisons. In an observational study, it may happen that an unmeasured bias that promotes A also promotes B, and in this case the differential effect may be less biased, or even unbiased, when other factorial contrasts are severely biased.

Whether the differential effect is of interest will depend upon the context. Obviously, the differential effect of A versus B is not the effect of A versus not-A; it may happen that A and B have substantial but similar effects, so their differential effect is small; in this case, the differential effect of A versus B reveals little about the effect of A. In other cases, the differential effect of A versus B may reveal quite a bit about the effect A versus not-A.

Examples of Differential Effects in Observational Studies

The study by Leonard Evans [16] in Sect. 1.4 is a skillful and fairly compelling use of differential effects to disambiguate treatment effects and unmeasured biases. Table 1.1 describes fatal accidents in which precisely one person in the front seat wore seat belts and precisely one of these two people died. There are, of course, many fatal accidents in which both people in the front seat wore seat belts and many others in which neither person wore seat belts. In other words, Table 1.1 describes the differential comparison, belted-driver-with-unbelted-passenger versus unbelted-driver-with-belted-passenger, which comprises two cells of a 2×2 factorial. As discussed in Sect. 1.4, the worry is that wearing seat belts is a precaution that may be accompanied by other precautions, driving at slower speeds or driving at a greater distance from the car ahead. A comparison that used the other cells of the 2×2 factorial—twice belted or twice unbelted—might be extremely biased, with use of seat belts taking the credit for cautious driving that occurred in unmeasured ways. In contrast, Table 1.1 compares two people in the same car, in the same accident, one belted, the other not, with the unbelted person at greater risk regardless of position. In a strict sense, Table 1.1 refers to the differential effect of belting with an unbelted companion, and that effect might be different from, say, belting with a belted companion: an unbelted companion may be hurled about the car and become a hazard to others. And yet, in this context, the beneficial differential effect of belting is strongly suggestive of a beneficial main effect of belting, and the differential effect is plausibly less affected by unmeasured biases.

Does regular use of nonsteroidal anti-inflammatory drugs (NSAIDs), such as the pain reliever ibuprofen, reduce the risk of Alzheimer disease? Several studies had found a negative association, but in a review of the subject, in 't Veld et al. [74] raised the possibility that people who are in the early stages of cognitive impairment might be less aware of pain or less active in seeking treatment for pain, thereby depressing their use of pain relievers. That is, perhaps the early stages of Alzheimer disease cause a reduction in the use of NSAIDs, rather than NSAIDs reducing the risk of Alzheimer disease. James Anthony et al. [7] addressed this possibility in the following way. Acetaminophen is a pain reliever that is not an NSAID. Anthony et al. [7, Table 2] found that the risk of Alzheimer disease among users of just NSAID pain relievers was about half that of users of just analgesic compounds that were not NSAIDs. A generic tendency to be less aware of pain could explain a negative association between Alzheimer disease and NSAIDs, but it does not explain the differential association that a preference for NSAIDs over other pain relievers is negatively associated with Alzheimer disease. Plausibly, Alzheimer disease causes people to be passive in the face of pain, but it is far less plausible that Alzheimer disease causes people to reach for acetaminophen rather than ibuprofen. The differential effect of acetaminophen versus ibuprofen is not plausibly affected by certain biases that plausibly affect the main effect of ibuprofen.

In practical cases it is not possible to demonstrate that differential effects are less affected than main effects by certain unobserved biases, precisely because the biases are not observed. It is possible, in an impractical way, to illustrate this phenomenon

Table 5.6 Attributes of the mothers who used marijuana, crack, neither or both in the U.S. National Institute on Drug Abuse's study of live births in the District of Columbia

	Neither	Marijuana	Crack	Both
Number of babies	931	11	39	5
Cigarette use during pregnancy (%)	22	100	92	100
Alcohol use during pregnancy (%)	23	64	74	100
Late or no prenatal care (%)	10	27	48	40
Less than high school education (%)	23	54	49	60
Married (%)	34	0	3	0
Did not want to get pregnant (%)	32	46	59	80

Mothers who used either marijuana or crack but not both are more similar to each other than either group is to mothers who used neither drug

by pretending for a moment that certain observed covariates are not observed. Table 5.6 compares four groups of mothers in the National Institute on Drug Abuse [107] study of babies born alive in eight hospitals in Washington, DC. None of these mothers reported using heroin during pregnancy, and the mothers are classified by whether they used marijuana, crack, neither or both during pregnancy. In terms of several variables, such as cigarette and alcohol use and prenatal care, mothers who used either drug are more similar to each other than they are to mothers who used neither drug. The differential effect of marijuana versus cocaine on the health of newborn babies may or may not be of interest; that depends upon the context. What Table 5.6 suggests is that this differential effect is likely to be less biased than either main effect formed by comparing with mothers who abstained from narcotics; see [135] for further discussion.

What Are Generic Biases?

A "generic unobserved bias" is one that promotes the use of both treatments [135]. The following definition from [135] makes this precise.

Definition 5.1 There is only generic unobserved bias in the assignment of two treatments, A and B, if

$$\frac{\Pr\left(Z_{A\ell}=1,\ Z_{B\ell}=0 \mid r_{T\ell},\ r_{C\ell},\ \mathbf{x}_\ell,\ u_\ell\right)}{\Pr\left(Z_{A\ell}=0,\ Z_{B\ell}=1 \mid r_{T\ell},\ r_{C\ell},\ \mathbf{x}_\ell,\ u_\ell\right)} = \vartheta\left(\mathbf{x}_\ell\right), \quad \ell = 1, \ldots, L, \tag{5.2}$$

for some (typically unknown) function $\vartheta\left(\mathbf{x}_\ell\right)$.

Condition (5.2) says that the odds of receiving treatment A in lieu of treatment B may depend on observed covariates, $\mathbf{x}_\ell$, but not on $(r_{T\ell},\ r_{C\ell},\ u_\ell)$. Condition (5.2) is considerably weaker than the naïve model's condition (3.5). That is, (5.2) may be true when quantities such as $\Pr\left(Z_{A\ell}=1 \mid r_{T\ell},\ r_{C\ell},\ \mathbf{x}_\ell,\ u_\ell\right)$ do depend upon $(r_{T\ell},\ r_{C\ell},\ u_\ell)$. For instance, in [7], (5.2) says that the odds of taking acetaminophen

rather than ibuprofen does not depend upon the unobservable $(r_{T\ell}, r_{C\ell}, u_\ell)$, but the chance of taking ibuprofen at all may depend upon $(r_{T\ell}, r_{C\ell}, u_\ell)$. In brief, (5.2) may be true when the naïve model is false.

It is possible to show [135] that a generic bias of the form (5.2) does not bias the differential effect of treatment A versus treatment B, although it may bias the main effects of A and of B. Specifically, suppose (5.2) is true with $0 < \vartheta(\mathbf{x}_\ell) < \infty$, and I matched pairs are formed, matching exactly for $\mathbf{x}_\ell$, in such a way that each pair contains one subject who received treatment A but not B and one subject who received treatment B but not A. Then the distribution of assignments to A in lieu of B within these matched pairs is the randomization distribution in Chap. 2; see [135] for proof. If (5.2) is false, then it is possible to conduct a sensitivity analysis that is concerned only with the degree of violation of (5.2) and not with other violations of the naïve model, and this sensitivity analysis closely resembles the sensitivity analysis in Chap. 3. See [135] for technical discussion, and [139] and [142, Chapter 12] for examples.

5.3 Instruments

5.3.1 What Is an Instrument?

An instrument is a random nudge to accept a treatment, where the nudge may or may not induce acceptance of the treatment, and the nudge can affect the outcome only if it does succeed in inducing acceptance of the treatment. Paul Holland [65] offered the "encouragement design" as a prototype of an instrument; see also the important paper by Joshua Angrist et al. [6]. In a paired, randomized encouragement design, individuals are paired based on measured pretreatment covariates, and for each pair, a coin is flipped to decide which member of the pair will be encouraged to accept the treatment. In an encouragement design, the experimenter hopes that encouraged individuals will accept the treatment and individuals who are not encouraged will decline the treatment; however, some individuals may in fact decline the treatment despite encouragement, and others may take the treatment in the absence of encouragement. A typical context would be encouragement to diet or exercise, or other similar settings in which a voluntary change in behavior is essential to the treatment.

There is, however, more to an instrument than randomly assigned encouragement. The encouragement itself must have no effect on the outcome besides its effect on acceptance of the treatment. This is often called the "exclusion restriction." Consider encouragement to exercise with weight loss as an outcome. To be an instrument, randomized encouragement to exercise must affect weight loss only by inducing exercise. Suppose you are indeed randomly selected and strongly encouraged to exercise. At the end of each week, you are told yet again that you are a failure and embarrassment for not exercising during the week, but despite this,

you will not exercise; instead, with all that good "encouragement," you fall into a deep depression, stop eating, and your weight drops. In this case, encouragement to exercise is not an instrument for exercise: it changed your weight, but not by inducing you to exercise.

Violations of the exclusion restriction are no small matter. If violation of the exclusion restriction is ignored in the exercise example, your weight loss will be attributed to exercise, though it was produced by depression instead.

Instruments are rare, but valuable when they exist.

5.3.2 Example: Noncompliance in a Double-Blind Randomized Trial

Randomized encouragement to accept either an active drug or a double-blind placebo is the experimental design that most closely approximates an instrument. In this case, encouragement is actually randomized. Also, because of the double-blind placebo, neither the subject nor the investigator knows what the subject is being encouraged to do. In this case, there are few opportunities for encouragement to affect a clinical outcome without shifting the amount of active drug consumed; that is, the exclusion restriction is likely to be satisfied. Even this situation is imperfect to the extent that side effects of the active drug alert the investigator or the subject to the identity of the treatment.

An example is from Jeffrey Silber et al.'s [153] randomized, double-blind clinical trial of the drug enalapril to protect the hearts of children who had received an anthracycline as part of cancer chemotherapy. Briefly, this is the "AAA trial" for the "ACE-inhibitor after anthracycline randomized trial." Anthracyclines are fairly effective at curing certain childhood cancers, but they may damage a child's heart. The study focused on children, teens mostly, whose cancer appeared to have been cured, but who showed signs of cardiac decline. The AAA trial randomized 135 children to enalapril or a double-blind placebo and measured cardiac function over several years. Although the children were encouraged to take a specific dose of drug or placebo, some children took less than that dose. It is likely that neither the child, nor the child's parents, nor the attending physician knew whether the child was taking less than a full dose of enalapril or less than a full dose of placebo [153], so it is likely that any effect of encouragement on cardiac function is a consequence of biological effects of the quantity of enalapril consumed. In this context, randomized encouragement is an instrument for the amount of enalapril actually consumed. Because it is an instrument, it provides a principled solution to the problem of noncompliance, that is, a principled answer to the question: How effective is the prescribed dose of drug on subjects willing to take that dose, despite the refusal of some subjects to take the prescribed dose? An instrumental variable analysis of noncompliance in the AAA trial is discussed by Robert Greevy et al. [57].

5.3.3 Example: Maimonides' Rule

Recall from Sects. 1.3 and 5.2.3 the study by Angrist and Lavy [4] of academic test performance and class size manipulated by Maimonides' rule. In that study, small variations in the size of the fifth grade cohort often produced large variations in class size for cohorts whose size was close to the cutpoint of 40 students; see Fig. 1.1. In Fig. 1.1, some schools had class sizes in defiance of Maimonides' rule; there is some noncompliance. Focus on schools with cohort sizes near 40, that is, the schools in Fig. 1.1. In this case, small variations in cohort size are presumably haphazard, while class size, not cohort size, is presumably what affects academic performance. If these two presumptions were actually correct for the schools in Fig. 1.1, then whether or not the cohort size exceeds 40 is an instrument for class size. In Table 5.2, in pair #3, there is noncompliance: in violation of Maimonides' rule, both schools in this pair had two small classes.

5.3.4 Notation for an Instrument in a Paired Encouragement Design

In a paired, randomized encouragement design, one person in pair i is picked at random and encouraged, denoted $Z_{ij} = 1$; the other person in the pair is not encouraged, denoted $Z_{ij} = 0$, so $Z_{i1} + Z_{i2} = 1$. The jth subject in pair i has two potential doses of treatment, dose d_{Tij} if encouraged to accept the treatment, $Z_{ij} = 1$, or dose d_{Cij} if not encouraged to accept the treatment, $Z_{ij} = 0$. We observed either d_{Tij} if $Z_{ij} = 1$ or d_{Cij} if $Z_{ij} = 0$, but the pair $\left(d_{Tij}, d_{Cij}\right)$ is not jointly observed for any subject ij. The dose of treatment actually received is d_{Tij} if $Z_{ij} = 1$ or d_{Cij} if $Z_{ij} = 0$, so it is $D_{ij} = Z_{ij}\, d_{Tij} + \left(1 - Z_{ij}\right)\, d_{Cij}$ in either case. The doses may be continuous, ordinal, or binary.[9] In Angrist and Lavy's [4] study, D_{ij} is the observed average class size, so $D_{11} = 23$ and $D_{12} = 40$ for the first pair in Table 5.2.

In Silber's AAA trial [57, 153], while on-study, a child assigned to placebo, $Z_{ij} = 0$, took no enalapril, $d_{Cij} = 0$, and a compliant child assigned to enalapril, $Z_{ij} = 1$, took the full assigned dose of enalapril, $d_{Tij} = 1$; however, many children were somewhat noncompliant and took somewhat less than the full dose, so $d_{Tij} < 1$ for many children ij. The decision by the experimenter to assign a child to enalapril or placebo, Z_{ij}, is randomized. The decision by the child to be compliant, $\left(d_{Tij} = 1, d_{Cij} = 0\right)$, or not compliant, $\left(d_{Tij} < 1, d_{Cij} = 0\right)$, is the child's decision, doubtless influenced by the child's parents. Quite possi-

[9] As in Note 8, when $\mathscr{F}$ was defined in Chap. 2, the potential doses $\left(d_{Tij}, d_{Cij}\right)$ were always equal to $(1, 0)$ and so were not mentioned. In general, if there are potential doses, $\left(d_{Tij}, d_{Cij}\right)$, then they are part of $\mathscr{F}$. Because previous discussions involving $\mathscr{F}$ had a single dose, we may adopt the new definition that includes doses in $\mathscr{F}$ without altering the content of those previous discussions.

bly, compliant children, $\left(d_{Tij} = 1, d_{Cij} = 0\right)$, differ from noncompliant children, $\left(d_{Tij} < 1, d_{Cij} = 0\right)$. Most of the children in the experiment were teens, newly engaging or rejecting sports, alcohol, tobacco, and narcotics. Perhaps compliant and noncompliant children are equally likely to smoke or engage in binge drinking, but perhaps not. Randomization has ensured that encouraged children are similar to unencouraged children, that the assignment to enalapril or placebo is equitable, but there is nothing to ensure that compliant children are comparable to less compliant children.

The paired randomized experiment in Chap. 2 is the special case in which encouragement is always perfectly decisive for a binary dose. Specifically, the paired randomized experiment is the special case with $\left(d_{Tij}, d_{Cij}\right) = (1, 0)$ for all i and j, so $D_{ij} = Z_{ij}$. In other words, people take the full dose of treatment, $D_{ij} = 1$, if encouraged to do so, $Z_{ij} = 1$, and they take none of the treatment, $D_{ij} = 0$, if not encouraged to accept the treatment, $Z_{ij} = 0$.

An instrument is said to be "strong" if d_{Tij} is much larger than d_{Cij} for most or all individuals ij; in this case, encouragement strongly shifts the dose received by most subjects. Figure 1.1 depicts a strong instrument: in many schools, average class size shifted decisively downward when enrollment passed 40 students. An instrument is said to be "weak" if d_{Tij} is close to or equal to d_{Cij} for most or all individuals ij; in this case, most individuals ignore encouragement. The most popular method of analysis using an instrument, namely two-stage least squares, tends to give incorrect inferences when an instrument is weak [28]: confidence intervals that claim 95% coverage may cover only 85% of the time. The confidence intervals and tests using an instrument in this book do not have this problem [71]: 95% intervals cover 95% of the time with weak and strong instruments.

An "intention-to-treat analysis" compares the encouraged group, $Z_{ij} = 1$, to the unencouraged group, $Z_{ij} = 0$, ignoring compliance behavior. It compares the intended treated group to the intended control group, ignoring actual treatment. The advantage of such an analysis is that it is fully justified by the random assignment of encouragement, Z_{ij}; that is, the groups being compared are comparable. The disadvantage is that "intention-to-treat analysis" estimates the effect of encouragement to accept the treatment, not the effect of the treatment itself. Both effects are interesting, but they can be quite different. Quitting smoking might be highly beneficial to your health, but difficult to do. Encouragement to quit smoking might be highly ineffective, because most people do not comply by quitting, but quitting itself might be highly effective. Both effects are important—the effects of encouragement to quit, the effects of quitting—but they are different things. In Silber's AAA trial [57, 153], if enalapril is beneficial, then enalapril consumed is likely to be more effective than enalapril in the bottle, and the intention-to-treat analysis estimates an idiosyncratic mixture of these two effects. Despite its limitations, and because of its strengths, the intention-to-treat analysis is one of the basic analyses that should be reported in any randomized trial with noncompliance. The analyses in Table 5.3 in Sect. 5.2.3 were intention-to-treat analyses: they ignored actual class size, D_{ij}, which sometimes violated Maimonides' rule. Is there a principled way to take account of noncompliance?

5.3.5 The Hypothesis that Effect Is Proportional to Dose

One hypothesis that satisfies the exclusion restriction asserts that the effect of encouragement, Z_{ij}, on the response, $\left(r_{Tij}, r_{Cij}\right)$, is proportional to its effect on the dose, $\left(d_{Tij}, d_{Cij}\right)$:

$$r_{Tij} - r_{Cij} = \beta \left(d_{Tij} - d_{Cij}\right) \quad \text{for} \quad i = 1, \ldots, I, \quad j = 1, 2. \tag{5.3}$$

In (5.3), if encouragement Z_{ij} does not affect your dose, $d_{Tij} = d_{Cij}$, then it does not affect your response, $r_{Tij} = r_{Cij}$. For instance, in Silber's AAA trial [57, 153], (5.3) implies that if assignment to enalapril, $Z_{ij} = 1$, does not actually induce a child to take enalapril, so $d_{Tij} = d_{Cij} = 0$, then it does not affect cardiac function, $r_{Tij} - r_{Cij} = 0$. In Silber's AAA trial, a compliant child has $\left(d_{Tij}, d_{Cij}\right) = (1, 0)$, so (5.3) would imply that such a child has $r_{Tij} - r_{Cij} = \beta\,(1 - 0) = \beta$, and β is the effect of the full dose, $d_{Tij} = 1$, for a compliant child. Similarly, for a less than fully compliant child who would take half the assigned dose, $d_{Tij} = \frac{1}{2}$, the hypothesis (5.3) implies an effect on cardiac function of $r_{Tij} - r_{Cij} = \beta\left(\frac{1}{2} - 0\right) = \beta/2$.

In Angrist and Lavy's [4] study, hypothesis (5.3) asserts that cohort size matters for test performance $\left(r_{Tij}, r_{Cij}\right)$ in direct proportion to its impact on class size, $\left(d_{Tij}, d_{Cij}\right)$. Expressed differently, we expect passing the cutpoint of 40 students to matter in pair $i = 1$ of Table 5.2, but not much in pair $i = 3$. In (5.3), β is the point gain in test performance, $r_{Tij} - r_{Cij}$, caused by a change in class size of $d_{Tij} - d_{Cij}$, and Fig. 1.1 suggests β is negative, with a decrease in class size increasing performance.

5.3.6 Inference About β

In an elementary but useful way, (5.3) may be rearranged as

$$r_{Tij} - \beta d_{Tij} = r_{Cij} - \beta d_{Cij} = a_{ij}, \quad \text{say}, \quad \text{for} \quad i = 1, \ldots, I, \quad j = 1, 2. \tag{5.4}$$

Suppose that we wish to test the hypothesis $H_0 : \beta = \beta_0$ in (5.3). From this hypothesis and the observable data, we may calculate the adjusted response, $R_{ij} - \beta_0 D_{ij}$. If the hypothesis $H_0 : \beta = \beta_0$ were true, then $R_{ij} - \beta_0 D_{ij} = r_{Tij} - \beta d_{Tij}$ if the jth subject in pair i is encouraged, $Z_{ij} = 1$, and $R_{ij} - \beta_0 D_{ij} = r_{Cij} - \beta d_{Cij}$ if this subject is not encouraged, so that in either case $R_{ij} - \beta_0 D_{ij} = a_{ij}$ using (5.4).[10] That is, if $H_0 : r_{Tij} - r_{Cij} = \beta_0 \left(d_{Tij} - d_{Cij}\right)$, $i = 1, \ldots, I$, $j = 1, 2$, were true, the adjusted responses, $R_{ij} - \beta_0 D_{ij} = a_{ij}$, would

[10] Recall Note 9. If the hypothesis $H_0 : r_{Tij} - r_{Cij} = \beta_0 \left(d_{Tij} - d_{Cij}\right)$ is true, then $R_{ij} - \beta_0 D_{ij} = a_{ij}$ is fixed, not varying with Z_{ij}. In other words, because the $\left(r_{Tij}, r_{Cij}, d_{Tij}, d_{Cij}\right)$'s are part of $\mathscr{F}$, if H_0 is true, then $\beta_0 = \left(r_{Tij} - r_{Cij}\right) / \left(d_{Tij} - d_{Cij}\right)$ is determined by $\mathscr{F}$, so using (5.4), the quantity a_{ij} may be calculated from $\mathscr{F}$.

satisfy the null hypothesis of no treatment effect. In light of this, the hypothesis $H_0 : \beta = \beta_0$ in (5.3) may be tested in a randomized encouragement design by calculating a statistic, such as Wilcoxon's signed rank statistic T, from the adjusted responses, and comparing T with its usual randomization distribution in Sect. 2.3.3; see [123]. Specifically, the signed rank statistic is computed from the encouraged ($Z_{ij} = 1$) minus unencouraged ($Z_{ij} = 0$) difference in adjusted responses, $Y_i^{(\beta_0)} = (Z_{i1} - Z_{i2})\{(R_{i1} - \beta_0 D_{i1}) - (R_{i2} - \beta_0 D_{i2})\}$. If $H_0 : \beta = \beta_0$ in (5.3) is true, then $Y_i^{(\beta_0)} = (Z_{i1} - Z_{i2})(a_{i1} - a_{i2})$ is $\pm|a_{i1} - a_{i2}|$; moreover, if $H_0 : \beta = \beta_0$ is true and encouragement, Z_{ij}, is randomized, then $Y_i^{(\beta_0)}$ is $a_{i1} - a_{i2}$ or $-(a_{i1} - a_{i2})$ each with probability $\frac{1}{2}$, so that, in particular, $Y_i^{(\beta_0)}$ is symmetrically distributed about zero. Notice that this reasoning is exactly parallel to Sect. 2.4; only the form of the hypothesis has changed [57, 71, 123, 126, 155].

The hypothesis $H_0 : \beta = 0$ in (5.3) is the hypothesis $H_0 : r_{Tij} = r_{Cij}$ for $i = 1, \ldots, I$, $j = 1, 2$; that is, $H_0 : \beta = 0$ in (5.3) is Fisher's sharp null hypothesis of no treatment effect. To reject either hypothesis is to reject the other: it is the same hypothesis, tested using the same test statistic, computed in the same way, compared with the same null distribution, yielding identical P-values. Because these two hypotheses are the same, the intention-to-treat analysis will reject Fisher's hypothesis of no treatment effect if and only if the instrumental variable (IV) analysis rejects $H_0 : \beta = 0$. As developed here, the IV analysis cannot find a treatment effect if the intention-to-treat analysis does not find one; indeed, they yield identical significance levels in testing the hypothesis of no effect.

Also, in strict parallel with Sect. 2.4, a $1 - \alpha$ confidence set for β is obtained by testing every hypothesis $H_0 : \beta = \beta_0$ and retaining for the confidence set the values not rejected by the test at level α; then, a $1 - \alpha$ confidence interval is the shortest interval containing the confidence set. This confidence set has the correct coverage rate even if the instrument is weak [71, 123]. If the instrument is extremely weak, the interval may compensate by becoming longer, and in extreme cases it may become infinite in length. A long confidence interval is a warning that the instrument is providing little information about the effect caused by the treatment. Better warned than misled.

Finally, in strict parallel with Sect. 2.4, a point estimate $\widehat{\beta}$ of β may be obtained by the method of Hodges and Lehmann; that is, by equating T computed from $R_{ij} - \beta_0 D_{ij}$ to its null expectation, $I(I+1)/4$, and solving the equation for the estimate, $\widehat{\beta}$.

The sensitivity of IV inferences to departures from randomly assigned encouragement applies the method of Sect. 3.4 to T when computed from the adjusted responses $R_{ij} - \beta_0 D_{ij}$; see [123, 126, 155]. Weak instruments are invariably sensitive to small departures from random assignment of encouragement [155].

5.3.7 Example: IV Analysis for Maimonides' Rule

In Angrist and Lavy's [4] study of academic test performance and class size manipulated by Maimonides' rule, the hypothesis $H_0 : \beta = \beta_0$ in (5.3) with $\beta_0 = -0.1$ for the math test asserts that a one-student increase in class size produces a decrease of $\frac{1}{10}$ point in the average math score. At the cutpoint of 40 students, Maimonides' rule reduces the average class size by about 20 students, so if $H_0 : \beta = -0.1$ were true, that reduction would increase average math test scores by about 2 points. Is that hypothesis plausible?

To test this hypothesis, apply Wilcoxon's signed rank test to

$$Y_i^{(\beta_0)} = (Z_{i1} - Z_{i2}) \{(R_{i1} - \beta_0 D_{i1}) - (R_{i2} - \beta_0 D_{i2})\} \tag{5.5}$$

with $\beta_0 = -0.1$. For instance, in Table 5.2, in pair $i = 1$,

$$Y_1^{(-0.1)} = (Z_{11} - Z_{12}) \{(R_{11} - \beta_0 D_{11}) - (R_{12} - \beta_0 D_{12})\} \tag{5.6}$$

$$= (1 - 0) [\{72.1 - (-0.1) \cdot 23.0\} - \{63.1 - (-0.1) \cdot 40.0\}] = 7.3. \tag{5.7}$$

Applying Wilcoxon's signed rank test to the $I = 86$ adjusted differences, $Y_i^{(-0.1)}$, yields a one-sided P-value of 0.00909, suggesting that $\beta < \beta_0 = -0.1$, and yields a two-sided P-value of $2 \times 0.00909 = 0.01818$. In Fig. 1.1, if cohort size were an instrument for class size—that is, if cohort size were random and influenced math scores only by influencing class size through (5.3)—then it would be quite clear that a one-student increase in class size depresses average math scores by more than $\frac{1}{10}$ of a point.

A confidence interval is built by inverting the hypothesis test; see Sect. 2.4. Testing every hypothesis $H_0 : \beta = \beta_0$ in (5.3) in this way, retaining those not rejected by the two-sided 0.05 level test, yields the 95% confidence interval for β of $[-0.812, -0.151]$. For instance, the one-sided P-value for testing $H_0 : \beta = -0.1515$ is 0.0252, so this hypothesis is barely not rejected in a two-sided, 0.05 level test, whereas the one-sided P-value for testing $H_0 : \beta = -1.51$ is 0.0249, so this hypothesis is barely rejected. Under Maimonides' rule, for a 20 student increase in class size as the cohort size goes from 41 to 40, the confidence interval for the change in average math scores 20β is $20 \times [-0.812, -0.151]$ or $[-16.24, -3.02]$ points.

A point estimate is obtained from the test by the method of Hodges and Lehmann; see Sect. 2.4. With $I = 86$ pairs of schools, the null expectation of Wilcoxon's signed rank statistic T is $I(I+1)/4 = 86(86+1)/4 = 1870.5$. If T is computed from $Y_i^{(-0.4518)}$, it yields $T = 1871$, a little too high, but if T is computed from $Y_i^{(-0.4519)}$, it yields $T = 1870$, a little too low, so upon rounding to three digits, the Hodges–Lehmann point estimate $\widehat{\beta}$ of β in (5.3) is $\widehat{\beta} = -0.452$ or $20\widehat{\beta} = -9.04$ math score points for a 20-student increase in class size.

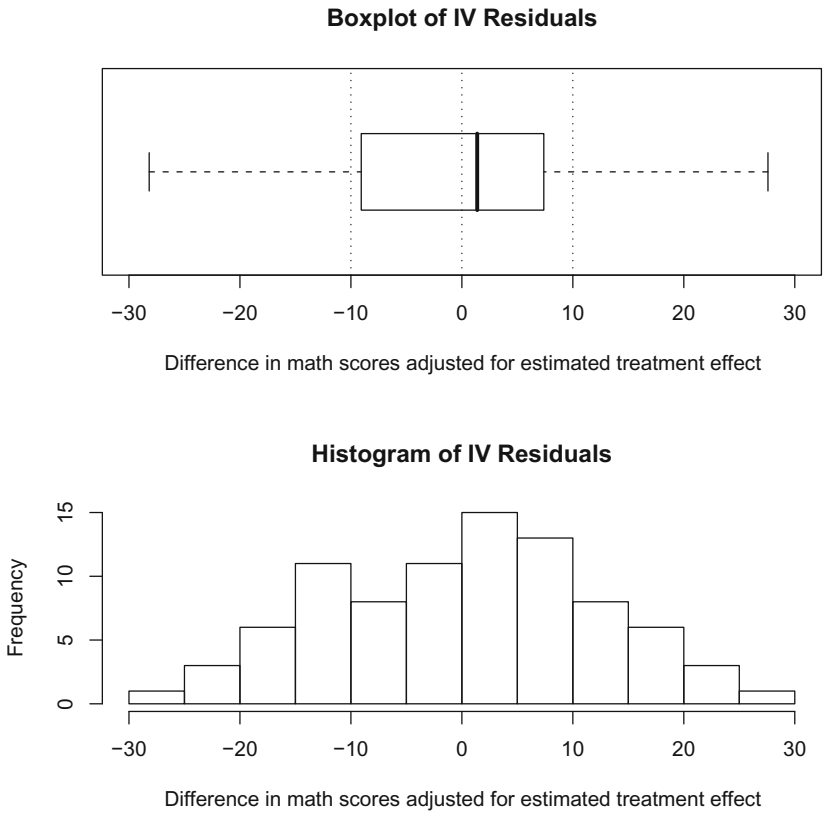

Fig. 5.2 IV residuals of math test scores, $Y_i^{(\beta_0)}$ with $\beta_0 = -0.452$, in $I = 86$ pairs of two Israeli schools. If cohort size were in fact an instrument for class size in these schools, and if $\beta = -0.452$, then these residuals would be symmetric about zero. The residuals look approximately symmetric about zero, with a mean of -0.03

If cohort size were truly an instrument for class size and if (5.3) were true as well, then $Y_i^{(\beta)}$ would be symmetric about zero. Figure 5.2 plots the "residuals," that is the differences in test performance adjusted for the estimated effect of class size, $Y_i^{(\widehat{\beta})} = Y_i^{(-0.452)}$. The residuals $Y_i^{(\widehat{\beta})}$ do not deviate noticeably from symmetry about zero.

As discussed in Sect. 5.2.3, for the 86 pairs of schools in Fig. 1.1 with fifth grade cohorts between 30 and 50, cohort size may not be completely random, so that cohort size may not actually be an instrument for class size. Because of the nature of the comparison, very large biases may seem implausible, but no matter how carefully a nonrandomized comparison is structured, we cannot rule out the possibility of small biases. As in Sects. 3.4 and 5.2.3, a sensitivity analysis may be conducted; see [123, 126] and [129, §5]. In fact, the math test results in Table 5.3 are the sensitivity analysis for the hypothesis $H_0 : \beta = 0$. For $\Gamma = 1.1$ or

$\Gamma = 1.2$, the maximum one-sided P-values for testing $H_0 : \beta = -0.1$ are, respectively, 0.0238 and 0.0508. For $\Gamma \geq 1$, there is an interval $\left[\widehat{\beta}_{\min}, \widehat{\beta}_{\max}\right]$ of possible point estimates of β where, as we have seen, for $\Gamma = 1$, the single estimate is $\widehat{\beta} = \widehat{\beta}_{\min} = \widehat{\beta}_{\max} = -0.452$. As Γ increases, $\widehat{\beta}_{\min}$ decreases and $\widehat{\beta}_{\max}$ increases. For $\Gamma = 1.83$, $\widehat{\beta}_{\max} = -0.1$, whereas for $\Gamma = 2.21$, $\widehat{\beta}_{\max} = 0$. In brief, from Table 5.3, rejection of no effect of class size is insensitive to small biases, $\Gamma \leq 1.4$, and a point estimate of no effect, $\widehat{\beta}_{\max} = 0$, requires $\Gamma = 2.21$. Again, the qualitative conclusions reached by Angrist and Lavy [4] are not extremely fragile; small deviations from a truly random instrument would not alter their qualitative conclusions.

5.3.8 Effect Ratios

Hypothesis (5.3) asserts that the effect of encouragement Z_{ij} on the outcome $r_{Tij} - r_{Cij}$ is proportional to its effect on the dose $d_{Tij} - d_{Cij}$. Is this hypothesis essential to the use of an instrument?

Proportional effects (5.3) are not essential. The situation is parallel to that in Sect. 2.4.4 where inference about an additive constant treatment effect τ was contrasted with inference about an average treatment effect ϑ. There, the model of a constant effect τ permitted the use of robust statistical procedures, and those robust procedures were not misled by treatment effects that varied from pair to pair but were symmetric about τ. However, Sect. 2.4.4 also noted that if the outcomes, (r_{Tij}, r_{Cij}), are short-tailed, then inference about the average treatment effect ϑ could be based on the randomization distribution of the average pair difference in outcomes, dispensing with the model of an additive constant treatment effect τ. As illustrated in Sect. 2.4.4, the price of shifting attention from τ to ϑ is a loss of robustness: both the estimator and the estimand, ϑ, may become unstable as $I \to \infty$ when the outcomes are long-tailed.

As discussed by Mike Baiocchi et al. [14, §3.2], an effect ratio is simply the ratio of two average treatment effects. For instance, in (5.8), the effect ratio λ is the average effect of encouragement on the outcome, (r_{Tij}, r_{Cij}), divided by the average effect of encouragement on the dose, (d_{Tij}, d_{Cij}),

$$\lambda = \frac{\sum_{i=1}^{I} \sum_{j=1}^{2} (r_{Tij} - r_{Cij})}{\sum_{i=1}^{I} \sum_{j=1}^{2} (d_{Tij} - d_{Cij})}. \tag{5.8}$$

As is true of the average treatment effect ϑ in Sect. 2.4.4, the effect ratio λ in (5.8) cannot be calculated because $r_{Tij} - r_{Cij}$ and $d_{Tij} - d_{Cij}$ are not observed. If effect were proportional to dose—if (5.3) were true—then substitution of (5.3) in (5.8) yields $\lambda = \beta$; however, λ in (5.8) is well-defined when (5.3) is not true. Importantly, an effect ratio (5.8) is well-defined whether or not the exclusion restriction holds.

Like the average treatment effect ϑ in Sect. 2.4.4, an effect ratio (5.8) may be unstable as $I \to \infty$ if (r_{Tij}, r_{Cij}) or (d_{Tij}, d_{Cij}) has long tails. Additionally, as $I \to \infty$ an effect ratio may be unstable despite short tails if $(2I)^{-1} \sum_{i=1}^{I} \sum_{j=1}^{2} (d_{Tij} - d_{Cij})$ is converging to zero. Recall that if an instrument is weak, then there is only a small effect of encouragement on the dose (d_{Tij}, d_{Cij}).

Consider testing the null hypothesis, $H_0 : \lambda = \lambda_0$, that the effect ratio, λ in (5.8), takes a specific value λ_0 in a paired experiment that has randomly assigned encouragement Z_{ij} in each pair i. In testing $H_0 : \lambda = \lambda_0$, we do not assume that the effect is proportional to dose (5.3). Use the specified value λ_0 in the hypothesis to define a new outcome, (r^*_{Tij}, r^*_{Cij}) as $r^*_{Tij} = r_{Tij} - \lambda_0 d_{Tij}$ and $r^*_{Cij} = r_{Cij} - \lambda_0 d_{Cij}$, and define ϑ^* to be the average effect of encouragement on this new outcome, so $\vartheta^* = (2I)^{-1} \sum_{i=1}^{I} \sum_{j=1}^{2} (r^*_{Tij} - r^*_{Cij})$. If $H_0 : \lambda = \lambda_0$ is true, then rearranging (5.8) yields $\vartheta^* = 0$. So, testing $H_0 : \lambda = \lambda_0$ is the same as testing the hypothesis that the new outcome (r^*_{Tij}, r^*_{Cij}) has zero average treatment effect [14, Proposition 2], a task accomplished in Sect. 2.4.4 for short-tailed outcomes. The test uses the randomization distribution of the mean pair difference, $\overline{Y}$, as in Sect. 2.4.4; however, this mean is computed from $Y_i^{(\lambda_0)}$ in (5.5). For a related sensitivity analysis for $\Gamma > 1$, see [46].

As in Sect. 2.4.4, this test of $H_0 : \lambda = \lambda_0$ is not robust, as it requires short-tailed outcomes and doses to permit the use of an average $\overline{Y}$; moreover, it will perform poorly if $(2I)^{-1} \sum_{i=1}^{I} \sum_{j=1}^{2} (d_{Tij} - d_{Cij})$ is near zero. Use of robust procedures to draw inferences about β in (5.3) will often be preferable [71]. Even with short tails, $\overline{Y}$ is not efficient in this context [43].

Inference about an effect ratio, λ in (5.8), does not require the exclusion restriction. Without the exclusion restriction, λ is simply the ratio of two average treatment effects. An important result of Angrist et al. [6] gives new meaning to λ when the exclusion restriction does hold. They assume that the dose, (d_{Tij}, d_{Cij}) is binary and monotone, $d_{Tij} \geq d_{Cij}$, with the consequence that (d_{Tij}, d_{Cij}) is $(1, 1)$, $(1, 0)$ or $(0, 0)$. If $(d_{Tij}, d_{Cij}) = (1, 0)$, then individual (i, j) is called a complier, because this individual complies with encouragement, with $D_{ij} = 1$ if encouraged, $Z_{ij} = 1$, or with $D_{ij} = 0$ if not encouraged, $Z_{ij} = 0$. They also assume encouragement has some effect on the dose, $0 \neq (2I)^{-1} \sum_{i=1}^{I} \sum_{j=1}^{2} (d_{Tij} - d_{Cij})$. Finally, they assume the exclusion restriction: $r_{Tij} - r_{Cij} = 0$ whenever $d_{Tij} - d_{Cij} = 0$. Under these assumptions, the denominator of λ in (5.8), namely $\sum_{i=1}^{I} \sum_{j=1}^{2} (d_{Tij} - d_{Cij})$, is the number of compliers. Terms in the numerator in (5.8) are zero except possibly for compliers, because of the exclusion restriction. So, λ in (5.8) becomes the average effect of the treatment on the subpopulation of compliers, or the complier-average-causal effect. This is surprising: we cannot identify compliers from the observed data, because we do not see (d_{Tij}, d_{Cij}). So, it is surprising that we can estimate the average effect of a treatment on a subpopulation that we cannot identify.

5.3.9 Is the Validity of an Instrument Testable?

To claim that one has an instrument is to assert several strong assumptions, including the exclusion restriction and the absence of bias in the assignment Z_{ij} of encouragement. Are these assumptions testable?

The bare structure, (R_{ij}, D_{ij}, Z_{ij}), $i = 1, \ldots, I$, $j = 1, 2$ offers few opportunities for testing these assumptions; however, in most scientific contexts, we have more than this bare structure. For instance, we may know that the treatment has negligible effects on some outcomes or on some individuals, as in Sect. 5.2.4. If we knew this, then one check on the instrument entails using the instrument to estimate an effect that we know, to see if it produces an estimate consistent with what we know.

For instance, there is debate about whether delivery by cesarean section improves the survival of extremely premature babies of 23–24 weeks gestational age. Delivery by cesarean section is thought to have little or no effect on the survival of healthy babies born at term, 39 weeks gestation age. Some hospitals do many cesarean sections, other hospitals do only a few. Is the rate of cesarean deliveries for term babies an instrument for cesarean delivery of extremely premature babies? To assert that it is an instrument is to assert several strong assumptions. For instance, hospitals that do many cesarean deliveries may be better or worse than other hospitals in several ways, not just in their use of cesarean section, so the exclusion restriction may be false. However, the belief that delivery by cesarean section is of little or no benefit to the survival of healthy babies born at term provides a check on these assumptions. In one application, the assumptions appeared to be implausible, because use of this supposed instrument found a large benefit of cesarean delivery both for premature babies, where such an effect is possible, and for babies born at term, where such an effect is not plausible; see Yang et al. [177]. For a different example, see Keele et al. [83].

5.3.10 When and Why Are Instruments Valuable?

At first glance, the use of instruments seems unattractive. A natural experiment seeks a consequential treatment to which people are exposed in a haphazard way. In contrast, an instrument requires both a haphazard exposure and additionally the exclusion restriction—that is, the exposure must affect the outcome only by affecting the treatment. If a natural experiment rests on one doubtful assumption, a study using an instrument rests on two doubtful assumptions. At first glance, the step towards instruments seems to be a step in the wrong direction.

This first impression is reinforced by a second. In reading the scientific literature, many claimed instruments seem implausible as instruments, perhaps because they are clearly not random, perhaps because the exclusion restriction seems implausible, or perhaps because the authors treat these central issues in a glib and superficial

manner. The examples from Angrist and Lavy [4] and Silber and colleagues [57, 153] are atypical in that they are at least plausible as instruments.

Given these two impressions, why seek instruments? When are they particularly valuable? When are they not needed? Instruments are of greatest value when any straightforward comparison of treated and control subjects is almost inevitably affected by the same bias, a bias that will recur whenever and wherever a straightforward comparison is made [71, §1]. A typical example is the effect of additional education on earnings. Almost inevitably, students who obtain post-secondary degrees have performed well in school during the ages of compulsory education, have greater motivation, better standardized test scores, typically greater financial resources, and so on, than students who leave school at an earlier stage. Repeatedly finding that people with more education earn more does little to isolate the effects actually caused by education; that is, the biases replicate with the same consistency that—perhaps with greater consistency than—the effects replicate [127]. In such a situation, it may be possible to find haphazard nudges that, at the margin, enable or discourage additional education, such as discontinuities in admission or aid criteria, proximity to affordable education, shifts in the cost or availability of aid. These nudges may be biased in various ways, but there may be no reason for them to be consistently biased in the same direction, so similar estimates of effect from studies subject to different potential biases gradually reduce ambiguity about what part is effect and what part is bias [127]. In this vein, David Card [33] surveys the many varied techniques used to estimate the economic returns actually caused by additional education.

5.4 Strengthening Weak Instruments

5.4.1 Why Strengthen an Instrument?

A weak instrument is a randomized but gentle encouragement to accept the treatment, encouragement that rarely changes the treatment an individual receives. In the language of Sect. 5.3 and [6], there are relatively few compliers. With a weak instrument, most people ignore encouragement; they do what they would have done anyway.

As discussed in Sect. 5.3 and [28, 43, 155], weak instruments create various problems. In particular, even with a perfectly valid instrument, the popular method of two-stage least squares gives mistaken answers with a weak instrument, too often reporting confidence intervals that fail to cover the true value of the parameter [28]; however, this particular problem is fixed by switching to better methods of estimation [71]. An instrument may not be perfectly valid, and it is commonplace, if not quite inevitable, that critics of a study will mention this. This criticism has force when the instrument is weak. Weak instruments are invariably sensitive to tiny imperfections, tiny departures from random assignment of encouragement [155];

that is, they are sensitive to Γ only slightly greater than 1. Even with a large treatment effect—a large β in (5.3)—a weak instrument means few people in the study are affected by this large treatment effect, with the consequence that even a large study has a small effective sample size, a problem made much worse by our inability to recognize and focus on the small subpopulation of compliers [43].

Is it possible to avoid these problems by strengthening an instrument?

5.4.2 A Popular Instrument in Health Outcomes Research: Distance to a Hospital

A central problem in nonrandomized studies of health outcomes is that the treatment a patient receives is adjusted to suit the health of that patient. If we saw poor outcomes following intensive treatment, then we would be uncertain about what to make of it. Does intensive treatment cause worse outcomes? Or is intensive treatment given to the sickest patients? Failure to completely record the pretreatment health of a patient may severely bias estimates of treatment effects. What appears to be harm caused by intensive treatment may instead reflect a decision to reserve intensive treatment for the most severely ill patients.

Is there an instrument?

In an emergency, people often seek care at a nearby hospital. If you have an acute myocardial infarction, or if you fall and break your hip, then you are unlikely to travel a long distance to reach a preferred hospital. If some hospitals have the facilities to provide a particular form of intensive treatment while other hospitals do not, then the care you receive in an emergency may be affected by the type of care available nearby. Some hospitals can implant a stent, others cannot. Some hospitals mostly use general anesthesia for hip fracture surgery, others mostly use regional anesthesia. Some hospitals have an advanced, high-level neonatal intensive care unit (NICU), others have a more limited NICU. Considerations of this kind have led to the use of geography as an instrument for particular types of health care that is available at some hospitals and not others [92, 99, 105, 106].

Does a hospital with a high-level NICU produce better outcomes for premature newborns than a hospital with a low-level NICU? Suppose that we were careful to compare patients who are similar in terms of demographic and socioeconomic covariates that vary from one neighborhood to another. Suppose that we measure the excess travel time to a hospital that has, say, a high-level NICU [14, 92]. Excess travel time is the travel time to the nearest hospital with a high-level NICU minus the travel time to the nearest hospital without a high-level NICU, and it can be negative if the nearest hospital has a high-level NICU. Is excess travel time an instrument for delivery at a hospital without a high-level NICU? A woman in labor, delivering prematurely, will think twice about traveling an extra hour to reach a more capable hospital, so sufficiently large excess travel time encourages delivery at a hospital without a high-level NICU. There is more to an instrument than encouragement to

accept a treatment. Aside from the demographic characteristics that were carefully controlled, living far from a high-level NICU has to be essentially random, that is, unrelated to the potential outcomes of the newborn. Perhaps that is plausible, or perhaps that would be plausible enough if small biases measured by Γ are insufficient to produce the false appearance of an effect.

5.4.3 How Is an Instrument Strengthened?

Does the use of regional anesthesia produce better outcomes than the use of general anesthesia during hip fracture surgery? In New York State, some hospitals commonly use regional anesthesia for hip fracture surgery, other hospitals commonly use general anesthesia, still other hospitals use regional anesthesia for some patients and general anesthesia for other patients. Some patients are geographically close to a hospital that prefers regional anesthesia, other patients are close to a hospital that prefers general anesthesia. We can easily look a hospital's past or historic preference for one type of anesthesia or another, and we can easily determine excess travel time to hospitals with different past preferences. Is a hospital's historic preference for one type of anesthesia an instrument for the anesthesia given to its current patients?

Mark Neuman et al. [105] present a map of New York State with hospitals color-coded by their historic preference for one type of anesthesia or another. The map is as you might anticipate. Manhattan needs to pulled from the state map and enlarged: otherwise, its hospitals are too numerous and too close together to be seen as separate entities. Everyone in Manhattan is geographically close to many hospitals with different preferences for anesthesia: excess travel times are near zero. Travel time to a hospital is barely relevant to hospital choice in Manhattan. There are cities in upstate New York, but on the map outside a few cities, hospitals are scattered about, in different colors without discernable pattern, at substantial distances from one another. Outside of cities, in the sparsely populated parts of upstate New York, many people are geographically close to precisely one hospital, so excess travel time is large in absolute value. In the sparse parts of upstate New York, excess travel time exerts a powerful push to receive the type of anesthesia that the local hospital happens to prefer. The instrument, excess travel time, is weak in Manhattan and strong in the sparsely populated parts of upstate New York.

How do you create a strong instrument? It is easy [14]. If you are looking for an art museum, elegant shopping, a fine restaurant, then go to Manhattan. If you are looking for a strong instrument, forget Manhattan; go to the sparsely populated parts of upstate New York. Do not insist upon using all of New York State, so that the cities with their many people and their weak instrument dominate the rural areas with their strong instrument. If you work in an office of the Governor of New York State and you must issue a report about all of New York State, then perhaps you have no choice; however, if you are a scientist looking for a natural experiment that provides compelling evidence about a treatment effect, then pick the best natural

experiment you can find. Neuman et al. [105] also eliminated hospitals with no decisive historical preference for one type of anesthesia.

Is Manhattan really as bad as all that? No, it is worse. What type of anesthesia will you receive for emergency hip fracture surgery? To some extent, it depends upon the preference of your hospital. What decides which hospital will perform your surgery? In the sparse parts of upstate New York, by and large, geography decides. In Manhattan, something else decides, who know what. Treatment assignment by "something else, who knows what" is not a good plan for a natural experiment. Remember, a strong instrument may be insensitive to small biases in the assignment of encouragement, but a weak instrument is invalidated by tiny biases [155]. Tisch Hospital is a teaching hospital of New York University Medical School and it is popular with residents of Manhattan who are free to choose their own hospital. Adjacent to Tisch Hospital, with zero excess travel time, is Bellevue Hospital, one of New York City's public hospitals. Bellevue Hospital often serves people with fewer choices; for instance, with its prison ward, Bellevue is a popular destination for the City's injured prisoners. Different parts of rural, upstate New York are not identical, but they differ less than the patient populations at Tisch and Bellevue hospitals. Where the instrument is strong, the instrument decides; where it is weak, something else decides, who knows what.

What is the tradeoff between a smaller sample size and a stronger instrument? Use all of New York State and you have an enormous sample size. Use the parts of New York State that afford a strong instrument and you have a large sample size. Is the loss of sample size worth it? Indeed it is. Ashkan Ertefaie et al. [43] perform detailed calculations. Consider a moderately large treatment effect and a flawless instrument entirely free of bias; then, a strong instrument with 50% compliers is slightly more efficient than a sample 25 times larger but with a weak instrument with 10% compliers [43, Table 2, $\Gamma = 1$, $\beta - \beta_0 = 0.5$]. In the same situation, if you want to report findings insensitive to a small bias of $\Gamma = 1.1$, then a strong instrument with 50% compliers is slightly more efficient than a sample 1300 times larger but with 10% compliers [43, Table 2, $\Gamma = 1.1$, $\beta - \beta_0 = 0.5$]. If you want to report insensitivity to a bias no longer small, $\Gamma = 1.25$, or a bias of moderate size, $\Gamma = 1.5$, then this is impossible with 10% compliers no matter how large the sample becomes; however, it is possible with 50% compliers.[11] If the strength of the instrument varies, then you are throwing away what is scarce and valuable—the strength of the instrument—by using all of the data, that is, by including the parts of the data in which the instrument is exceedingly weak.

[11] Part III of this book develops the concept of design sensitivity, the limiting sensitivity to bias as the sample size increases, $I \to \infty$. In [43, Table 2, $\beta - \beta_0 = 0.5$], the design sensitivity is $\widetilde{\Gamma} = 1.73$ with 50% compliers, or $\widetilde{\Gamma} = 1.11$ with 10% compliers. In this specific situation, with 10% compliers, results will be sensitive to a bias $\Gamma > 1.11$ in sufficiently large samples. These calculations assume that Wilcoxon's signed rank statistic is the basis for the test. The quoted results from [43] make use of design sensitivity and the Bahadur efficiency of a sensitivity analysis with an instrument. See Sect. 19.5 and [140] for discussion of the Bahadur efficiency of a sensitivity analysis.

5.4.4 Use of Matching in Strengthening an Instrument

In the example in the previous section, excess travel time is near zero in certain regions of New York State, so its strength as an instrument is increased by excluding data from those regions. New York State provided a simple story, a narrative, about instrument strength in relation to the structure of the state. However, a narrative is not needed to strengthen an instrument.

Mike Baiocchi et al. [14, 15] strengthened an instrument using a matching technique, nonbipartite matching, that is described in Chap. 12. The traditional but awkward phrase "nonbipartite" means simply "not two parts," or not treatment-control. In nonbipartite matching, there is a single group of individuals and a distance defined between every two individuals in the group. An optimal nonbipartite match forms nonoverlapping pairs of two distinct individuals to minimize the total distance within the pairs. Nonbipartite matching is described in detail in Chap. 12 and [95].

In the study of neonatal intensive care units (NICUs) described earlier in this section [14, 92], Baiocchi and colleagues used excess travel time to a hospital in Pennsylvania with a high-level NICU as an instrument discouraging delivery at such a hospital. A distance was defined between preterm babies born in Pennsylvania between 1995 and 2004. The distance was small if the two babies had similar covariates and very different excess travel times. The distance was large if either the babies were substantially different in terms of covariates or had similar excess travel times. An optimal nonbipartite match formed pairs of babies who looked similar in terms of covariates, but for whom excess travel time encouraged delivery at hospitals with different levels of NICU.

Covariates were variables determined prior to the baby's delivery. These included demographic and socioeconomic covariates, such as mother's age, race, parity, education, health insurance and neighborhood attributes from the US Census. In Pennsylvania, demographic covariates vary geographically, so it is important to control these when geography is the instrument. The covariates also included birth weight and gestational age, two important predictors of perinatal outcomes.

As an illustration of instrument strengthening, Baiocchi and colleagues built two matched samples from the same data, a larger sample with a weaker instrument, and a smaller sample with a stronger instrument. The larger sample had 99,174 pairs of babies. The smaller sample had 49,587 pairs of babies. In the larger but weaker sample, the longer-minus-shorter difference in travel time was, on average, 13.5 min; that is, the trip to the hospital with a high-level NICU took 13.5 min longer [14, Table 1]. In the smaller but stronger sample, the difference in travel time was, on average, 34.2 min, or about 2.5 times longer than in the larger sample. Again, consider what an extra 13.5 min, or an extra 34.2 min, means to a mother in labor with a preterm baby.

As might be expected, an extra 13.5 min was less of a deterrent than an extra 34.2 min. In each pair, one baby is nearer to a high-level NICU and the other further away. In the larger but weaker match, 35% of near babies and 53% of

far babies were delivered at a hospital without a high-level NICU [14, Table 2]. In the smaller but stronger match, 31% of near babies and 75% of far babies were delivered at a hospital without a high-level NICU. So, the instrument was strengthened not just in concept but in fact, not just in terms of travel time, but in delivery location. Excess travel time did more to determine delivery location in the smaller, stronger match. Strengthening an instrument means creating stronger encouragement, a stronger Z_{ij}, here a greater difference in excess travel time. We hope to see that stronger encouragement translates into a larger difference in dose (d_{Tij}, d_{Cij}), here a larger difference in delivery at a hospital with a high-level NICU. We cannot use the realized value D_{ij} of (d_{Tij}, d_{Cij}) in matching, because D_{ij} is an outcome of encouragement, Z_{ij}, but we can certainly strengthen the encouragement.

The two matches, large but weak, small but strong, produced similar estimates of the effect ratio (5.8), about a 0.9% reduction in mortality of preterm babies at hospitals with a high-level NICU, not quite 1 baby in 100 babies. The standard error of this estimate was quite different: the standard error was about 70% larger—yes, larger—with more babies than with fewer babies, because the instrument was so much stronger in the smaller, stronger match [14, Table 3]. This is an important empirical reminder of a theoretical fact [43]: the standard error is not a mirror of the sample size, particularly with weak instruments. More importantly, a small bias, $\Gamma = 1.07$, could explain away the ostensible effect of a high-level NICU in the larger, weaker match, but the smaller, stronger match was insensitive to a bias of $\Gamma = 1.22$; see [14, Tables 4–5]. Again, this pattern in sensitivity to unmeasured bias in the instrument is anticipated by theory [155].

5.5 Summary

Although structured by competing theories that make conflicting predictions, observational studies are built from opportunities, devices, and instruments.

Opportunities are special circumstances in which competing theories may be contrasted with unusual clarity. In these circumstances, the theories make sharply different predictions and the most common or plausible unmeasured biases are reduced or absent.

The association between a treatment and an outcome is ambiguous when treatments are not assigned at random. The association may be produced by an effect caused by the treatment or by a comparison of people who are not truly comparable. Devices are the tools of disambiguation, the active effort to reduce ambiguity.

Certain treatments are consistently assigned in the same biased way, so that straightforward comparisons of treated and control groups inevitably reproduce the same bias. There is no direct opportunity to separate the bias from the treatment effect. In these cases, it may be possible to find haphazard nudges towards treatment, pushes that tilt but do not control treatment assignment. To the extent that these nudges are assigned at random and affect outcomes only through acceptance of the

treatment, they form instruments that may permit the isolation of treatment effects. Instruments are rare, but useful when they exist.

5.6 Further Reading

Parts III and IV of this book and Chapters 6 through 9 of [129] are "further reading" for this chapter. Part III discusses "design sensitivity" which measures the degree to which different designs or data generating processes produce results that are insensitive to unmeasured biases. In contrast, Chapters 6 through 8 of [129] discuss the circumstances under which known effects or multiple control groups can detect unmeasured biases, presenting technical results about these topics, while Chapter 9 of [129] discusses some technical aspects of coherence; alternatively, for articles covering similar material, see [117–120, 124]. Multiple control groups are discussed in [19, 21, 31, 66, 79, 93, 109, 111, 117, 119, 143, 159, 178] and [129, §8]; see also Sects. 12.3 and 23.3 and Chap. 21. Known effects are discussed in [12, 29, 39, 90, 91, 100, 113, 114, 118, 119, 163, 172, 173], and [129, §6]. Differential effects and generic biases are discussed further in Chap. 13; see also [135, 181, 182] and [142, Chapter 12], where further results and examples may be found. A related, interesting but different design is used by Rahul Roychoudhuri et al. [146].

The view of instruments as a bridge between randomized experiments and observational studies is discussed in various ways in [3, 6, 14–17, 42, 43, 48, 53, 57, 69, 72, 76–78, 169]. The link between instruments and randomization inference is developed in [57, 71, 123, 126, 130]. The relationship between instrument strength and sensitivity to unmeasured biases in the instrument is discussed in [43, 155]. Matching to strengthen an instrument is discussed in [14, 15, 43, 179]; see also Chap. 12. Use of conventional bipartite matching to strengthen an instrument is discussed in [177, Appendix I].

"Opportunities" are discussed in [1, 3, 10, 13, 56, 60, 80, 88, 102, 125, 144, 147, 148, 161, 167, 175]. Lotteries often present "opportunities"; see, for instance, [5, 17, 64, 73].

5.7 Software

In the `DOS2` package in `R`, the function `cohere` performs the coherent signed rank test in Sect. 5.2.3 and [124]; see also Chap. 18.

Chapter 12 discusses software for nonbipartite matching, as used in Sect. 5.4. The `nbpMatching` package [95] in `R` performs optimal nonbipartite matching using the algorithm of Derigs [40].

5.8 Data

Part of Angrist and Lavy's [4] data in Fig. 1.1 is in the `DOS2` package in `R` as `angristlavy`. The data in Fig. 5.1 from Pinto et al. [112] are in `pinto` in the `DOS2` package.

Bibliography

1. Abadie, A., Cattaneo, M.D.: Econometric methods for program evaluation. Ann. Rev. Econ. **10**, 465–503 (2018)
2. Abadie, A., Gardeazabal, J.: Economic costs of conflict: a case study of the Basque Country. Am. Econ. Rev. **93**, 113–132 (2003)
3. Angrist, J.D., Krueger, A.B.: Empirical strategies in labor economics. In: Ashenfelter, O., Card, D. (eds.) Handbook of Labor Economics, vol. 3, pp. 1277–1366. Elsevier, New York (1999)
4. Angrist, J.D., Lavy, V.: Using Maimonides' rule to estimate the effect of class size on scholastic achievement. Q. J. Econ. **114**, 533–575 (1999)
5. Angrist, J., Lavy, V.: New evidence on classroom computers and pupil learning. Econ. J. **112**, 735–765 (2002)
6. Angrist, J.D., Imbens, G.W., Rubin, D.B.: Identification of causal effects using instrumental variables (with Discussion). J. Am. Stat. Assoc. **91**, 444–455 (1996)
7. Anthony, J.C., Breitner, J.C., Zandi, P.P., Meyer, M.R., Jurasova, I., Norton, M.C., Stone, S.V.: Reduced prevalence of AD in users of NSAIDs and H_2 receptor antagonists. Neurology **54**, 2066–2071 (2000)
8. Ares, M., Hernandez, E.: The corrosive effect of corruption on trust in politicians: evidence from a natural experiment. Res. Politics April–June, 1–8 (2017)
9. Armstrong, C.S.: Discussion of "CEO compensation and corporate risk-taking: evidence from a natural experiment." J. Account. Econ. **56**, 102–111 (2013)
10. Armstrong, C.S., Kepler, J.D.: Theory, research design assumptions, and causal inferences. J. Account. Econ. **66**, 366–373 (2018)
11. Armstrong, C.S., Blouin, J.L., Larcker, D.F.: The incentives for tax planning. J. Account. Econ. **53**, 391–411 (2012)
12. Arnold, B.F., Ercumen, A., Benjamin-Chung, J., Colford, J.M.: Negative controls to detect selection bias and measurement bias in epidemiologic studies. Epidemiology **27**, 637–641 (2016)
13. Athey, S., Imbens, G.W.: The state of applied econometrics: causality and policy evaluation. J. Econ. Perspect **31**, 3–32 (2018)
14. Baiocchi, M., Small, D.S., Lorch, S., Rosenbaum, P.R.: Building a stronger instrument in an observational study of perinatal care for premature infants. J. Am. Stat. Assoc. **105**, 1285–1296 (2010)
15. Baiocchi, M., Small, D.S., Yang, L., Polsky, D., Groeneveld, P.W.: Near/far matching: a study design approach to instrumental variables. Health Serv. Outcomes Res. Method **12**, 237–253 (2012)
16. Barnard, J., Du, J.T., Hill, J.L., Rubin, D.B.: A broader template for analyzing broken randomized experiments. Sociol. Methods Res. **27**, 285–317 (1998)
17. Barnard, J., Frangakis, C.E., Hill, J.L., Rubin, D.B.: Principal stratification approach to broken randomized experiments: a case study of School Choice vouchers in New York City. J. Am. Stat. Assoc. **98**, 299–311 (2003)

18. Basta, N.E., Halloran, M.E.: Evaluating the effectiveness of vaccines using a regression discontinuity design. Am. J. Epidemiol. **188**, 987–990 (2019)
19. Battistin, E., Rettore, E.: Ineligibles and eligible non-participants as a double comparison group in regression-discontinuity designs. J. Econometrics **142**, 715–730 (2008)
20. Beautrais, A.L., Gibb, S.J., Fergusson, D.M., Horwood, L.J., Larkin, G.L.: Removing bridge barriers stimulates suicides: an unfortunate natural experiment. Austral. New Zeal. J. Psychiatry **43**, 495–497 (2009)
21. Behrman, J.R., Cheng, Y., Todd, P.E.: Evaluating preschool programs when length of exposure to the program varies: a nonparametric approach. Rev. Econ. Stat. **86**, 108–132 (2004)
22. Berk, R.A., de Leeuw, J.: An evaluation of California's inmate classification system using a regression discontinuity design. J. Am. Stat. Assoc. **94**, 1045–1052 (1999)
23. Berk, R.A., Rauma, D.: Capitalizing on nonrandom assignment to treatments: a regression-discontinuity evaluation of a crime-control program. J. Am. Stat. Assoc. **78**, 21–27 (1983)
24. Bernanke, B.S.: The macroeconomics of the Great Depression: a comparative approach. J. Money Cred. Bank **27**, 1–28 (1995). Reprinted: Bernanke, B.S. Essays on the Great Depression. Princeton University Press, Princeton (2000)
25. Bilban, M., Jakopin, C.B.: Incidence of cytogenetic damage in lead-zinc mine workers exposed to radon. Mutagenesis **20**, 187–191 (2005)
26. Black, S.: Do better schools matter? Parental valuation of elementary education. Q. J. Econ. **114**, 577–599 (1999)
27. Bound, J.: The health and earnings of rejected disability insurance applicants. Am. Econ. Rev. **79**, 482–503 (1989)
28. Bound, J., Jaeger, D.A., Baker, R.M.: Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. J. Am. Stat. Assoc. **90**, 443–450 (1995)
29. Brew, B.K., Gong, T., Williams, D.M., Larsson, H., Almqvist, C.: Using fathers as a negative control exposure to test the Developmental Origins of Health and Disease Hypothesis: a case study on maternal distress and offspring asthma using Swedish register data. Scand. J. Public Health **45**(Suppl. 17), 36–40 (2017)
30. Campbell, D.T.: Factors relevant to the validity of experiments in social settings. Psychol. Bull. **54**, 297–312 (1957)
31. Campbell, D.T.: Prospective: artifact and control. In: Rosenthal, R., Rosnow, R. (eds.) Artifact in Behavioral Research, pp. 351–382. Academic, New York (1969)
32. Campbell, D.T.: Methodology and Epistemology for Social Science: Selected Papers. University of Chicago Press, Chicago (1988)
33. Card, D.: The causal effect of education. In: Ashenfelter, O., Card, D., (eds.) Handbook of Labor Economics. North Holland, New York (2001)
34. Choudhri, E.U., Kochin, L.A.: The exchange rate and the international transmission of business cycle disturbances: some evidence from the Great Depression. J. Money Cred. Bank **12**, 565–574 (1980)
35. Cioran, E.M.: History and Utopia. University of Chicago Press, Chicago (1998)
36. Cochran, W.G.: The planning of observational studies of human populations (with Discussion). J. R. Stat. Soc. A **128**, 234–265 (1965)
37. Conley, T.G., Hansen, C.B., Rossi, P.E.: Plausibly exogenous. Rev. Econ. Stat. **94**, 260–272 (2012)
38. Cook, T.D.: Waiting for life to arrive: a history of the regression-discontinuity designs in psychology, statistics and economics. J. Econometrics **142**, 636–654 (2007)
39. Davey Smith, G.: Negative control exposures in epidemiologic studies. Epidemiology **23**, 350–351 (2012)
40. Derigs, U.: Solving nonbipartite matching problems by shortest path techniques. Ann. Oper. Res. **13**, 225–261 (1988)
41. Eichengreen, B., Sachs, J.: Exchange rates and economic recovery in the 1930's. J. Econ. Hist. **45**, 925–946 (1985)

42. Ertefaie, A., Small, D.S., Flory, J.H., Hennessy, S.: A tutorial on the use of instrumental variables in pharmacoepidemiology. Pharmacoepidemiol. Drug Saf. **26**, 357–367 (2017)
43. Ertefaie, A., Small, D.S., Rosenbaum, P.R.: Quantitative evaluation of the trade-off of strengthened instruments and sample size in observational studies. J. Am. Stat. Assoc. **113**, 1122–1134 (2018)
44. Evans, L.: The effectiveness of safety belts in preventing fatalities. Accid. Anal. Prev. **18**, 229–241 (1986)
45. Fenech, M., Chang, W.P., Kirsch-Volders, M., Holland, N., Bonassi, S., Zeiger, E.: HUMN project: detailed description of the scoring criteria for the cytokinesis-block micronucleus assay using isolated human lymphocyte cultures. Mutat. Res. **534**, 65–75 (2003)
46. Fogarty, C.B.: Studentized sensitivity analysis for the sample average treatment effect in paired observational studies. J. Am. Stat. Assoc. (2019, to appear). https://doi.org/10.1080/01621459.2019.1632072
47. Fogarty, C.B., Small, D.S.: Sensitivity analysis for multiple comparisons in matched observational studies through quadratically constrained linear programming. J. Am. Stat. Assoc. **111**, 1820–1830 (2016)
48. Frangakis, C.E., Rubin, D.B.: Addressing complications of intention-to-treat analysis in the combined presence of all-or-none treatment noncompliance and subsequent missing outcomes. Biometrika **86**, 365–379 (1999)
49. French, B., Cologne, J., Sakata, R., Utada, M., Preston, D.L.: Selection of reference groups in the Life Span Study of atomic bomb survivors. Eur. J. Epidemiol. **32**, 1055–1063 (2017)
50. Friedman, M., Schwartz, A.J.: A Monetary History of the United States. Princeton University Press, Princeton (1963)
51. Frye, T., Yakovlev, A.: Elections and property rights: a natural experiment from Russia. Comp. Pol. Stud. **49**, 499–528 (2016)
52. Gangl, M.: Causal inference in sociological research. Ann. Rev. Sociol. **36**, 21–47 (2010)
53. Goetghebeur, E., Loeys, T.: Beyond intent to treat. Epidemiol. Rev. **24**, 85–90 (2002)
54. Gormley, T.A., Matsa, D.A., Milbourn, T.: CEO compensation and corporate risk-taking: evidence from a natural experiment. J. Account. Econ. **56**, 79–101 (2013)
55. Gould, E.D., Lavy, V., Paserman, M.D.: Immigrating to opportunity: estimating the effect of school quality using a natural experiment on Ethiopians in Israel. Q. J. Econ. **119**, 489–526 (2004)
56. Gow, I.D., Larcker, D.F., Reiss, P.C.: Causal inference in accounting research. J. Account. Res. **54**, 477–523 (2016)
57. Greevy, R., Silber, J.H., Cnaan, A., Rosenbaum, P.R.: Randomization inference with imperfect compliance in the ACE-inhibitor after anthracycline randomized trial. J. Am. Stat. Assoc. **99**, 7–15 (2004)
58. Guo, Z., Kang, H., Cai, T.T., Small, D.S.: Confidence interval for causal effects with invalid instruments using two-stage hard thresholding with voting. J. R. Stat. Soc. B **80**, 793–815 (2018)
59. Hahn, J., Todd, P., Van der Klaauw, W.: Identification and estimation of treatment effects with a regression-discontinuity design. Econometrica **69**, 201–209 (2001)
60. Hamermesh, D.S.: The craft of labormetrics. Ind. Labor Relat. Rev. **53**, 363–380 (2000)
61. Hawkins, N.G., Sanson-Fisher, R.W., Shakeshaft, A., D'Este, C., Green, L.W.: The multiple baseline design for evaluating population based research. Am. J. Prev. Med. **33**, 162–168 (2007)
62. Heckman, J., Navarro-Lozano, S.: Using matching, instrumental variables, and control functions to estimate economic choice models. Rev. Econ. Stat. **86**, 30–57 (2004)
63. Hill, A.B.: The environment and disease: association or causation? Proc. R. Soc. Med. **58**, 295–300 (1965)
64. Ho, D.E., Imai, K.: Estimating the causal effects of ballot order from a randomized natural experiment: California alphabet lottery, 1978–2002. Public Opin. Q. **72**, 216–240 (2008)
65. Holland, P.W.: Causal Inference, path analysis, and recursive structural equations models. Sociol. Method **18**, 449–484 (1988)

66. Holland, P.W.: Choosing among alternative nonexperimental methods for estimating the impact of social programs: comment. J. Am. Stat. Assoc. **84**, 875–877 (1989)
67. Imbens, G.W.: The role of the propensity score in estimating dose response functions. Biometrika **87**, 706–710 (2000)
68. Imbens, G.W.: Nonparametric estimation of average treatment effects under exogeneity: a review. Rev. Econ. Stat. **86**, 4–29 (2004)
69. Imbens, G.W.: Instrumental variables: an econometrician's perspective. Stat. Sci. **29**, 323–358 (2014)
70. Imbens, G.W., Lemieux, T.: Regression discontinuity designs: a guide to practice. J. Econometrics **142**, 615–635 (2008)
71. Imbens, G., Rosenbaum, P.R.: Robust, accurate confidence intervals with a weak instrument: quarter of birth and education. J. R. Stat. Soc. A **168**, 109–126 (2005)
72. Imbens, G.W., Rubin, D.B.: Causal Inference in Statistics, Social, and Biomedical Sciences. Cambridge University Press, New York (2015)
73. Imbens, G.W., Rubin, D.B., Sacerdote, B.I.: Estimating the effect of unearned income on labor earnings, savings, and consumption: evidence from a survey of lottery players. Am. Econ. Rev. **91**, 778–794 (2001)
74. in 't Veld, B.A., Launer, L.J., Breteler, M.M.B., Hofman, A., Stricker, B.H.C.: Pharmacologic agents associated with a preventive effect on Alzheimer's disease. Epidemiol. Rev. **2**, 248–268 (2002)
75. Joffe, M.M., Colditz, G.A.: Restriction as a method for reducing bias in the estimation of direct effects. Stat. Med. **17**, 2233–2249 (1998)
76. Kang, H.: Matched instrumental variables. Epidemiology **27**, 624–632 (2016)
77. Kang, H., Zhang, A., Cai, T.T., Small, D.S.: Instrumental variables estimation with some invalid instruments and its application to Mendelian randomization. J. Am. Stat. Assoc. **111**, 132–144 (2016)
78. Kang, H., Peck, L., Keele, L.: Inference for instrumental variables: a randomization inference approach. J. R. Stat. Soc. **181**, 1231–1254 (2018)
79. Karmakar, B., Small, D.S., Rosenbaum, P.R.: Using approximation algorithms to build evidence factors and related designs for observational studies. J. Comp. Graph. Stat. **28**(3), 698–709 (2019)
80. Keele, L.: The statistics of causal inference: a view from political methodology. Polit. Anal. **23**, 313–335 (2015)
81. Keele, L., Morgan, J.W.: How strong is strong enough? Strengthening instruments through matching and weak instrument tests. Ann. Appl. Stat. **10**, 1086–1106 (2016)
82. Keele, L., Titiunik, R., Zubizarreta, J.R.: Enhancing a geographic regression discontinuity design through matching to estimate the effect of ballot initiatives on voter turnout. J. R. Stat. Assoc. A **178**, 223–239 (2015)
83. Keele, L., Zhao, Q., Kelz, R.R., Small, D.S.: Falsification tests for instrumental variable designs with an application to the tendency to operate. Med. Care **57**, 167–171 (2019)
84. Keele, L., Harris, S., Grieve, R.: Does transfer to intensive care units reduce mortality? A comparison of an instrumental variables design to risk adjustment. Med. Care **57**, e73–e79 (2019)
85. Khuder, S.A., Milz, S., Jordan, T., Price, J., Silvestri, K., Butler, P.: The impact of a smoking ban on hospital admissions for coronary heart disease. Prev. Med. **45**, 3–8 (2007)
86. LaLumia, S.: The effects of joint taxation of married couples on labor supply and non-wage income. J. Public Econ. **92**, 1698–1719 (2008)
87. Lambe, M., Cummings, P.: The shift to and from daylight savings time and motor vehicle crashes. Accid. Anal. Prev. **32**, 609–611 (2002)
88. Lawlor, D.A., Tilling, K., Davey Smith, G.: Triangulation in aetiological epidemiology. Int. J. Epidemiol. **45**, 1866–1886 (2016)
89. Li, F., Frangakis, C.E.: Polydesigns and causal inference. Biometrics **62**, 343–351 (2006)

90. Liew, Z., Kioumourtzoglou, M.A., Roberts, A.L., O'Reilly, E.J., Ascherio, A., Weisskopf, M.G.: Use of negative control exposure analysis to evaluate confounding: an example of acetaminophen exposure and attention-deficit/hyperactivity disorder in Nurses' Health Study II. Am. J. Epidemiol. **188**, 768–775 (2019)
91. Lipsitch, M., Tchetgen Tchetgen, E.J., Cohen, T.: Negative controls: a tool for detecting confounding and bias in observational studies. Epidemiology **21**, 383–388 (2010)
92. Lorch, S.A., Baiocchi, M., Ahlberg, C.E., Small, D.S.: The differential impact of delivery hospital on the outcomes of premature infants. Pediatrics **130**, 270–278 (2012)
93. Lu, B., Rosenbaum, P.R.: Optimal matching with two control groups. J. Comput. Graph Stat. **13**, 422–434 (2004)
94. Lu, X., White, H.: Robustness checks and robustness tests in applied economics. J. Econometrics **178**, 194–206 (2014)
95. Lu, B., Greevy, R., Xu, X., Beck, C.: Optimal nonbipartite matching and its statistical applications. Am. Stat. **65**, 21–30 (2011)
96. Ludwig, J., Miller, D.L.: Does Head Start improve children's life chances? Evidence from a regression discontinuity design. Q. J. Econ. **122**, 159–208 (2007)
97. Manski, C.: Nonparametric bounds on treatment effects. Am. Econ. Rev. **80**, 319–323 (1990)
98. Marquart, J.W., Sorensen, J.R.: Institutional and postrelease behavior of Furman-commuted inmates in Texas. Criminology **26**, 677–693 (1988)
99. McClellan, M., McNeil, B.J., Newhouse, J.P.: Does more intensive treatment of acute myocardial infarction in the elderly reduce mortality? J. Am. Med. Assoc. **272**, 859–866 (1994)
100. McKillip, J.: Research without control groups: a control construct design. In: Bryant, F.B., et al. (eds.) Methodological Issues in Applied Social Psychology, pp. 159–175. Plenum Press, New York (1992)
101. Mealli, F., Rampichini, C.: Evaluating the effects of university grants by using regression discontinuity designs. J. R. Stat. Soc. A **175**, 775–798 (2012)
102. Meyer, B.D.: Natural and quasi-experiments in economics. J. Bus. Econ. Stat. **13**, 151–161 (1995)
103. Mill, J.S.: On Liberty. Barnes and Nobel, New York (1859, reprinted 2004)
104. Milyo, J., Waldfogel, J.: The effect of price advertising on prices: evidence in the wake of 44 Liquormart. Am. Econ. Rev. **89**, 1081–1096 (1999)
105. Neuman, M.D., Rosenbaum, P.R., Ludwig, J.M., Zubizarreta, J.R., Silber, J.H.: Anesthesia technique, mortality and length of stay after hip fracture surgery. J. Am. Med. Assoc. **311**, 2508–2517 (2014)
106. Newhouse, J.P., McClellan, M.: Econometrics in outcomes research: the use of instrumental variables. Ann. Rev. Public Health **19**, 17–34 (1998)
107. NIDA: Washington DC Metropolitan Area Drug Study (DC*MADS), 1992. U.S. National Institute on Drug Abuse: ICPSR Study No. 2347 (1999). http://www.icpsr.umich.edu
108. Oreopoulos, P.: Long-run consequences of living in a poor neighborhood. Q. J. Econ. **118**, 1533–1575 (2003)
109. Origo, F.: Flexible pay, firm performance and the role of unions: new evidence from Italy. Labour Econ. **16**, 64–78 (2009)
110. Peto, R., Pike, M., Armitage, P., Breslow, N., Cox, D., Howard, S., Mantel, N., McPherson, K., Peto, J., Smith, P.: Design and analysis of randomised clinical trials requiring prolonged observation of each patient, I. Br. J. Cancer **34**, 585–612 (1976)
111. Pimentel, S.D., Small, D.S., Rosenbaum, P.R.: Constructed second control groups and attenuation of unmeasured biases. J. Am. Stat. Assoc. **111**, 1157–1167 (2016)
112. Pinto, D., Ceballos, J.M., García, G., Guzmán, P., Del Razo, L.M., Gómez, E.V.H., García, A., Gonsebatt, M.E.: Increased cytogenetic damage in outdoor painters. Mutat. Res. **467**, 105–111 (2000)
113. Reynolds, K.D., West, S.G.: A multiplist strategy for strengthening nonequivalent control group designs. Eval. Rev. **11**, 691–714 (1987)

114. Rosenbaum, P.R.: From association to causation in observational studies. J. Am. Stat. Assoc. **79**, 41–48 (1984)
115. Rosenbaum, P.R.: The consequences of adjustment for a concomitant variable that has been affected by the treatment. J. R. Stat. Soc. A **147**, 656–666 (1984)
116. Rosenbaum, P.R.: Sensitivity analysis for certain permutation inferences in matched observational studies. Biometrika **74**, 13–26 (1987)
117. Rosenbaum, P.R.: The role of a second control group in an observational study (with Discussion). Stat. Sci. **2**, 292–316 (1987)
118. Rosenbaum, P.R.: The role of known effects in observational studies. Biometrics **45**, 557–569 (1989)
119. Rosenbaum, P.R.: On permutation tests for hidden biases in observational studies. Ann. Stat. **17**, 643–653 (1989)
120. Rosenbaum, P.R.: Some poset statistics. Ann. Stat. **19**, 1091–1097 (1991)
121. Rosenbaum, P.R.: Detecting bias with confidence in observational studies. Biometrika **79**, 367–374 (1992)
122. Rosenbaum, P.R.: Hodges-Lehmann point estimates in observational studies. J. Am. Stat. Assoc. **88**, 1250–1253 (1993)
123. Rosenbaum, P.R.: Comment on a paper by Angrist, Imbens, and Rubin. J. Am. Stat. Assoc. **91**, 465–468 (1996)
124. Rosenbaum, P.R.: Signed rank statistics for coherent predictions. Biometrics **53**, 556–566 (1997)
125. Rosenbaum, P.R.: Choice as an alternative to control in observational studies (with Discussion). Stat. Sci. **14**, 259–304 (1999)
126. Rosenbaum, P.R.: Using quantile averages in matched observational studies. Appl. Stat. **48**, 63–78 (1999)
127. Rosenbaum, P.R.: Replicating effects and biases. Am. Stat. **55**, 223–227 (2001)
128. Rosenbaum, P.R.: Stability in the absence of treatment. J. Am. Stat. Assoc. **96**, 210–219 (2001)
129. Rosenbaum, P.R.: Observational Studies, 2nd edn. Springer, New York (2002)
130. Rosenbaum, P.R.: Covariance adjustment in randomized experiments and observational studies (with Discussion). Stat. Sci. **17**, 286–327 (2002)
131. Rosenbaum, P.R.: Does a dose-response relationship reduce sensitivity to hidden bias? Biostatistics **4**, 1–10 (2003)
132. Rosenbaum, P.R.: Design sensitivity in observational studies. Biometrika **91**, 153–164 (2004)
133. Rosenbaum, P.R.: Heterogeneity and causality: unit heterogeneity and design sensitivity in observational studies. Am. Stat. **59**, 147–152 (2005)
134. Rosenbaum, P.R.: Exact, nonparametric inference when doses are measured with random errors. J. Am. Stat. Assoc. **100**, 511–518 (2005)
135. Rosenbaum, P.R.: Differential effects and generic biases in observational studies. Biometrika **93**, 573–586 (2006)
136. Rosenbaum, P.R.: What aspects of the design of an observational study affect its sensitivity to bias from covariates that were not observed? Festschrift for Paul W. Holland. ETS, Princeton (2009)
137. Rosenbaum, P.R.: Testing one hypothesis twice in observational studies. Biometrika **99**, 763–774 (2012)
138. Rosenbaum, P.R.: Nonreactive and purely reactive doses in observational studies. In: Berzuini, C., Dawid, A.P., Bernardinelli, L. (eds.) Causality: Statistical Perspectives and Applications, pp. 273–289. Wiley, New York (2012)
139. Rosenbaum, P.R.: Using differential comparisons in observational studies. Chance **26**(3), 18–23 (2013)
140. Rosenbaum, P.R.: Bahadur efficiency of sensitivity analyses in observational studies. J. Am. Stat. Assoc. **110**, 205–217 (2015)
141. Rosenbaum. P.R.: How to see more in observational studies: some new quasi-experimental devices. Ann. Rev. Stat. Appl. **2**, 21–48 (2015)

142. Rosenbaum, P.R.: Observation and Experiment: An Introduction to Causal Inference. Harvard University Press, Cambridge (2017)
143. Rosenbaum, P.R., Silber, J.H.: Using the exterior match to compare two entwined matched control groups. Am. Stat. **67**, 67–75 (2013)
144. Rosenzweig, M.R., Wolpin, K.I.: Natural 'natural experiments' in economics. J. Econ. Lit. **38**, 827–874 (2000)
145. Rothman, K.J.: Modern Epidemiology. Little, Brown, Boston (1986)
146. Roychoudhuri, R., Robinson, D., Putcha, V., Cuzick, J., Darby, S., Møller, H.: Increased cardiovascular mortality more than fifteen years after radiotherapy for breast cancer: a population-based study. BMC Cancer **7**, 9 (2007)
147. Rutter, M.: Proceeding from observed correlation to causal inference: the use of natural experiments. Perspect. Psychol. Sci. **2**, 377–395 (2007)
148. Rutter, M.: Identifying the Environmental Causes of Disease: How Do We Decide What to Believe and When to Take Action? Academy of Medical Sciences, London (2007)
149. Sekhon, J.S.: Opiates for the matches: matching methods for causal inference. Ann. Rev. Pol. Sci. **12**, 487–508 (2009)
150. Sekhon, J.S., Titiunik, R.: When natural experiments are neither natural nor experiments. Am. Pol. Sci. Rev. **106**, 35–57 (2012)
151. Sennett, R.: The Uses of Disorder. Yale University Press, New Haven (1971, 2008)
152. Shadish, W.R., Cook, T.D.: The renaissance of field experimentation in evaluating interventions. Annu. Rev. Psychol. **60**, 607–629 (2009)
153. Silber, J.H., Cnaan, A., Clark, B.J., Paridon, S.M., Chin, A.J., et al.: Enalapril to prevent cardiac function decline in long-term survivors of pediatric cancer exposed to anthracyclines. J. Clin. Oncol. **5**, 820–828 (2004)
154. Small, D.S.: Sensitivity analysis for instrumental variables regression with overidentifying restrictions. J. Am. Stat. Assoc. **102**, 1049–1058 (2007)
155. Small, D.S., Rosenbaum, P.R.: War and wages: the strength of instrumental variables and their sensitivity to unobserved biases. J. Am. Stat. Assoc. **103**, 924–933 (2008)
156. Small, D.S., Rosenbaum, P.R.: Error-free milestones in error-prone measurements. Ann. Appl. Stat. **3**, 881–901 (2009)
157. Sobel, M.E.: An introduction to causal inference. Sociol. Methods Res. **24**, 353–379 (1996)
158. Sommer, A., Zeger, S.L.: On estimating efficacy from clinical trials. Stat. Med. **10**, 45–52 (1991)
159. Stuart, E.A., Rubin, D.B.: Matching with multiple control groups with adjustment for group differences. J. Educ. Behav. Stat. **33**, 279–306 (2008)
160. Sullivan, J.M., Flannagan, M.J.: The role of ambient light level in fatal crashes: inferences from daylight saving time transitions. Accid. Anal. Prev. **34**, 487–498 (2002)
161. Summers, L.H.: The scientific illusion in empirical macroeconomics (with Discussion). Scand. J. Econ. **93**, 129–148 (1991)
162. Tan, Z.: Regression and weighting methods for causal inference using instrumental variables. J. Am. Stat. Assoc. **101**, 1607–1618 (2006)
163. Tchetgen Tchetgen, E.J.: The control outcome calibration approach for causal inference with unobserved confounding. Am. J. Epidemiol. **179**, 633–640 (2013)
164. Thistlethwaite, D.L., Campbell, D.T.: Regression-discontinuity analysis. J. Educ. Psychol. **51**, 309–317 (1960)
165. Trochim, W.M.K.: Pattern matching, validity and conceptualization in program evaluation. Eval. Rev. **9**, 575–604 (1985)
166. van Eeden, C.: An analogue, for signed rank statistics, of Jureckova's asymptotic linearity theorem for rank statistics. Ann. Math. Stat. **43**, 791–802 (1972)
167. Vandenbroucke, J.P.: When are observational studies as credible as randomized trials? Lancet **363**, 1728–1731 (2004)
168. Varian, H.R.: Causal inference in economics and marketing. Proc. Natl. Acad. Sci. **113**, 7310–7315 (2016)

169. Wang, X., Jiang, Y., Zhang, N.R., Small, D.S.: Sensitivity analysis and power for instrumental variable studies. Biometrics **74**, 1150–1160 (2018)
170. Weed, D.L., Hursting, S.D.: Biologic plausibility in causal inference: current method and practice. Am. J. Epidemiol. **147**, 415–425 (1998)
171. Weiss, N.: Inferring causal relationships: elaboration of the criterion of dose-response. Am. J. Epidemiol. **113**, 487–490 (1981)
172. Weiss, N.: Can the 'specificity' of an association be rehabilitated as a basis for supporting a causal hypothesis? Epidemiology **13**, 6–8 (2002)
173. West, S.G., Duan, N., Pequegnat, W., Gaist, P., Des Jarlais, D.C., Holtgrave, D., Szapocznik, J., Fishbein, M., Rapkin, B., Clatts, M., Mullen, P.D.: Alternatives to the randomized controlled trial. Am. J. Public Health **98**, 1359–1366 (2008)
174. Wintemute, G.J., Wright, M.A., Drake, C.M., Beaumont, J.J.: Subsequent criminal activity among violent misdemeanants who seek to purchase handguns: risk factors and effectiveness of denying handgun purchase. J. Am. Med. Assoc. **285**, 1019–1026 (2001)
175. Wolpin, K.I.: The Limits of Inference Without Theory. MIT Press, Cambridge (2013)
176. Wright, M.A., Wintemute, G.J., Rivara, F.P.: Effectiveness of denial of handgun purchase to persons believed to be at high risk for firearm violence. Am. J. Public Health **89**, 88–90 (1999)
177. Yang, F., Zubizarreta, J.R., Small, D.S., Lorch, S., Rosenbaum, P.R.: Dissonant conclusions when testing the validity of an instrumental variable. Am. Stat. **68**, 253–263 (2014)
178. Yoon, F.B., Huskamp, H.A., Busch, A.B., Normand, S.L.T.: Using multiple control groups and matching to address unobserved biases in comparative effectiveness research: an observational study of the effectiveness of mental health parity. Stat. Biosci. **3**, 63–78 (2011)
179. Zubizarreta, J.R., Small, D.S., Goyal, N.K., Lorch, S., Rosenbaum, P.R.: Stronger instruments via integer programming in an observational study of late preterm birth outcomes. Ann. App. Stat. **7**, 25–50 (2013)
180. Zubizarreta, J.R., Cerda, M., Rosenbaum, P.R.: Effect of the 2010 Chilean earthquake on posttraumatic stress: reducing sensitivity to unmeasured bias through study design. Epidemiology **7**, 79–87 (2013)
181. Zubizarreta, J.R., Small, D.S., Rosenbaum, P.R.: Isolation in the construction of natural experiments. Ann. Appl. Stat. **8**, 2096–2121 (2014)
182. Zubizarreta, J.R., Small, D.S., Rosenbaum, P.R.: A simple example of isolation in building a natural experiment. Chance **31**, 16–23 (2018)

Chapter 6
Transparency

Abstract Transparency means making evidence evident. An observational study that is not transparent may be overwhelming or intimidating, but it is unlikely to be convincing. Several aspects of transparency are briefly discussed.

> The beliefs we have most warrant for, have no safeguard to rest on, but a standing invitation to the whole world to prove them unfounded. If the challenge is not accepted, or is accepted and the attempt fails, we are far enough from certainty still; but we have done the best the existing state of human reason admits of. . . . This is the amount of certainty attainable by a fallible being, and this is the sole way of attaining it.
>
> John Stuart Mill [4, p. 21]

> The objectivity of all science, including mathematics, is inseparably linked with its criticizability.
>
> Karl R. Popper [6, p. 137]

Transparency means making evidence evident. An experiment, and by analogy an observational study, is not a private experience, not the source of some private conviction. In the absence of transparency, evidence, argument, and conclusions are not fully available for critical evaluation. Critical discussion, the standing invitation to undermine evidence or argument, is, for John Stuart Mill, the sole safeguard of our most warranted beliefs. For Karl Popper, critical discussion is inseparably linked with the objectivity of science and mathematics. To the extent that transparency is needed for critical discussion, transparency is no small issue.

David Cox [2, p. 8] writes:

> An important aspect of analysis, difficult to achieve with complex methods, is transparency. That is, in principle, the pathway between the data and the conclusions should be as clear as is feasible. This is partly for the self-education of the analyst and also is for protection against errors. . . . It is also important for presenting conclusions. . . . Transparency strongly encourages the use of the simplest methods that will be adequate.

See also Cox [1, p. 11].

In addition to the "use of the simplest methods that will be adequate," transparency is aided by the following considerations. Introducing a felicitous phrase,

© Springer Nature Switzerland AG 2020

P. R. Rosenbaum, *Design of Observational Studies*, Springer Series in Statistics,
https://doi.org/10.1007/978-3-030-46405-9_6

Mervyn Susser [8, p. 74] writes: "A main object of research design is thus to *simplify the conditions of observation*." In a chapter with that title, Susser [8, Chapter 7] discusses: (1) limiting observations to the relevant segment of the population to "isolate the hypothetical causal variable and permit study of its effects alone" [8, p. 74], (2) "selecting suitable situations" [8, p. 76], and (3) matching to remove bias from observed covariates. Concerning (1), see also Joffe and Colditz [3], who use the term "restriction." In discussing the design of clinical trials, Richard Peto et al. [5, p. 590] write: "A positive result is more likely, and a null result is more informative, if the main comparison is of only two treatments, these being as different as possible."

Modularity aids transparency. If a study addresses several issues, each potentially contentious, but the study is composed of separable, simple modular components, then there is the realistic prospect that the scope of contention will be constructively focused. Modularity limits leakage: controversy remains contained within the region of contention. Here are three modular questions that arise in most observational studies. Do the treated and untreated populations overlap sufficiently with respect to observed covariates to permit the construction of a comparable control group, or do the populations need to be restricted before such a comparison is attempted? Has matching succeeded in balancing observed covariates, so matched treated and control groups are comparable with respect to these observed covariates? Is it plausible that differing outcomes in treated and control groups are produced by imbalances in a specific unmeasured covariate? If these questions are separated and addressed one at a time, the first two questions may be settled with little controversy before engaging the inevitably more contentious third question. If the three questions are addressed simultaneously in one grand analysis, little conviction or consensus may develop even about the first two questions.

To Cox's good list of reasons for seeking transparency, I would add one more [7, Chapter 12]. If smaller issues are permitted to become unnecessarily complex, then the analysis may collapse under the weight of these complexities before engaging larger issues. In observational studies, if the adjustment for observed covariates becomes unnecessarily complex, as sometimes happens, then the analysis may never engage the fundamental issue, namely possible biases from covariates that were not measured. Like a good card trick, unnecessarily complex adjustments for observed covariates may distract attention from the fundamental issue in observational studies, but they are unlikely to shed much light on it.

An empirical study that is not transparent may be published or cited, but it is less likely to undergo serious critical discussion and therefore less likely to receive the implicit endorsement of surviving such discussion largely unscathed. Conversely, if critical discussion of a transparent study reveals potential ambiguities or alternative interpretations, that discussion may stimulate a replication that eliminates the ambiguities (Sect. 4.5). If a study has unambiguous faults, they are more likely to be discovered if the study is transparent, and hence there is greater reason to trust the conclusions of a transparent study when no such faults are identified.

References

1. Cox, D.R.: Planning of Experiments. Wiley, New York (1958)
2. Cox, D.R.: Applied statistics: a review. Ann. Appl. Stat. **1**, 1–16 (2007)
3. Joffe, M.M., Colditz, G.A.: Restriction as a method for reducing bias in the estimation of direct effects. Stat. Med. **17**, 2233–2249 (1998)
4. Mill, J.S.: On Liberty. Barnes and Nobel, New York (1859, reprinted 2004)
5. Peto, R., Pike, M., Armitage, P., Breslow, N., Cox, D., Howard, S., Mantel, N., McPherson, K., Peto, J., Smith, P.: Design and analysis of randomised clinical trials requiring prolonged observation of each patient, I. Br. J. Cancer **34**, 585–612 (1976)
6. Popper, K.R.: Objective Knowledge. Oxford University Press, Oxford (1972)
7. Rosenbaum, P.R.: Observational Studies. Springer, New York (2002)
8. Susser, M.: Causal Thinking in the Health Sciences. Oxford University Press, New York (1973)

Chapter 7
Some Counterclaims Undermine Themselves

Abstract A counterclaim disputes the claim that an association between treatment received and outcome exhibited reflects an effect caused by the treatment. Some counterclaims undermine themselves. A supplemental statistical analysis may demonstrate this.

7.1 Appraising Counterclaims

7.1.1 Types of Counterclaims

A symphony is followed by applause. An observational study is followed by objections. The mere fact that objections are raised is unremarkable and, by itself, uninformative. We must attend to the specifics: the quality of the objections; the degree to which they are specific, coherent, and plausible[1]; their compatibility with the data at hand.

Objections are counterclaims or plausible rival explanations of observed associations as something other than an effect caused by the treatment. Objections may be serious or frivolous, specific, or vague, public-spirited or self-serving, and sometimes difficult to locate on these dimensions. A cogent counterclaim may protect us from an unwarranted claim. It can strengthen an observational study if the only objections are frivolous, vague and self-serving: Is that the strongest criticism a critic can muster?

Irwin Bross [2] wrote:

> ... [S]ome critics appear to labor under the misconception that if some flaw can be found in a study, this automatically invalidates the author's conclusions ... [It] is not enough to spot flaws in a study: a responsible critic would go on to show how these flaws lead to a counterhypothesis that can explain the observations. ... Proponents of scientific hypotheses are often justly criticized for their "tubular vision"—a remarkable inability to see the

[1] Hill [5] invites consideration of these dimensions of a causal claim, but they also merit consideration when evaluating objections or counterclaims. See also [12, 13].

© Springer Nature Switzerland AG 2020

P. R. Rosenbaum, *Design of Observational Studies*, Springer Series in Statistics,
https://doi.org/10.1007/978-3-030-46405-9_7

> evidence unfavorable to their hypothesis. Critics are equally subject to this type of defective vision.

The philosopher J. L. Austin [1, pp. 84–87] raised the following objection to vague objections that demand ever more evidence:

> If you say "That's not enough," then you must have in mind some more or less definite lack ... Enough is enough: it doesn't mean everything ... The wile of the metaphysician consists in asking "Is it a real table?" (a kind of object which has no obvious way of being phony) and not specifying or limiting what may be wrong with it, so that I feel at a loss "how to prove" it is a real one.

Perhaps surprisingly, a specific, serious counterclaim can be an opportunity to strengthen the evidence in support of the original claim. Some counterclaims undermine themselves [10].

7.1.2 The Logic of Counterclaims That Undermine Themselves

How can a counterclaim undermine itself? A familiar, if extreme, example occurs in a proof by contradiction: supposing that proposition P is true is shown to lead to a contradiction, so we conclude that P is not true after all. Assuming P permits a deduction that would otherwise be impermissible, and that deduction ends in a contradiction. The premise P of a proof by contradiction leads to an unintelligible tangle best avoided.

In a statistical analysis, an asserted counterclaim P can undermine itself in the following, more limited, way. If the investigator does not believe P, then the investigator is in no position to suppose that P is true, and so is in no position to invoke P as an assumption underlying a particular statistical analysis. Left alone, the investigator would not perform any analysis that takes P as a premise. This situation changes when a critic asserts that P is true. If a critic asserts the counterclaim P, then the investigator might reasonably respond: "Well, I myself do not believe P, but let us assume P is true for the sake of argument. If we assume P, then the following additional statistical analysis is permissible that would otherwise be impermissible. So let us conduct that additional statistical analysis. Using the data, this additional statistical analysis that assumes P is true is seen to provide stronger evidence in support of the original claim than did the original statistical analysis that did not assume P. In this sense, assuming P would only strengthen the evidence, not weaken it."

When a counterclaim undermines itself, this merely shows the counterclaim fails in its role as a counterclaim. Every statistical analysis acknowledges that a certain level of uncertainty attends its conclusions: a confidence interval is not a single point, a small P-value could be due to extraordinary luck. When a counterclaim undermines itself, supposing the counterclaim to be true reduces the acknowledged level of uncertainty rather than increasing it. Importantly, a counterclaim that undermines itself does not justify a shorter confidence interval or a smaller P-value;

rather, a counterclaim that undermines itself shows that certain kinds of criticism of the original confidence interval or P-value are not viable criticisms. If ostensibly plausible criticisms are not viable as criticisms, then demonstrating this is a form of strengthening of the original claim.

In an observational study, this reasoning takes a particular form. The investigator performs an analysis suggestive of an effect caused by the treatment, acknowledging that this analysis is sensitive to unmeasured biases of a particular magnitude, Γ in Sect. 3.4. A critic asserts a counterclaim, P. The investigator observes that, were P true, an additional statistical analysis would be justified, and when that analysis is conducted, its conclusions in support of the original claim are insensitive to a larger bias, $\Gamma' > \Gamma$. In this case, were P true, the bias required to explain away the observed associations as not causal has become larger, not smaller; that is, uncertainty about unmeasured bias would be reduced, not increased. In this specific sense, if P undermines itself, then P is not a viable counterclaim to the evidence as originally presented.

7.2 An Example: Safety Belts, Injury, and Ejection

7.2.1 A First Look at an Example: A Claim Prior to a Counterclaim

Motivated by Evans' [3] study in Sects. 1.4 and 5.2.6 of the effects of safety belts in car crashes, Fig. 7.1 examines more recent data from the US Fatality Analysis Reporting System for 2010–2011, as discussed in detail in [10].[2] Figure 7.1 compares the injuries experienced by a driver and a front passenger in the same car in the same crash. Injuries are scored 0 for no injury, 1 for possible injury, 2 for non-incapacitating injury, 3 for incapacitating injury, and 4 for death. The figure shows the driver-minus-passenger difference in injury scores, −4 to 4, where −4 means that the driver was uninjured and the passenger died. Figure 7.1 distinguishes four situations defined by whether safety belts were used. In situation (B,B), both driver and passenger wore lap-shoulder safety belts. In situation (U,U), both driver and passenger were not belted. In situation (B,U), the driver wore a lap-shoulder belt and the passenger was unbelted. In situation (U,B), the driver was unbelted and the passenger wore a lap-shoulder belt. Some other unusual situations are excluded from Fig. 7.1. Very old cars may have a lap-belt without a shoulder belt, and this situation was excluded.

[2]Because the US Fatality Analysis Reporting System only records information about crashes with at least one fatality, it has certain limitations—certain "problems of ascertainment" [4]—that affect the types of inferences that may be drawn. These limitations do not affect tests of the hypothesis of no treatment effect, as developed in this chapter, but they would have consequences for estimates of the magnitude of effect that are not discussed here [10].

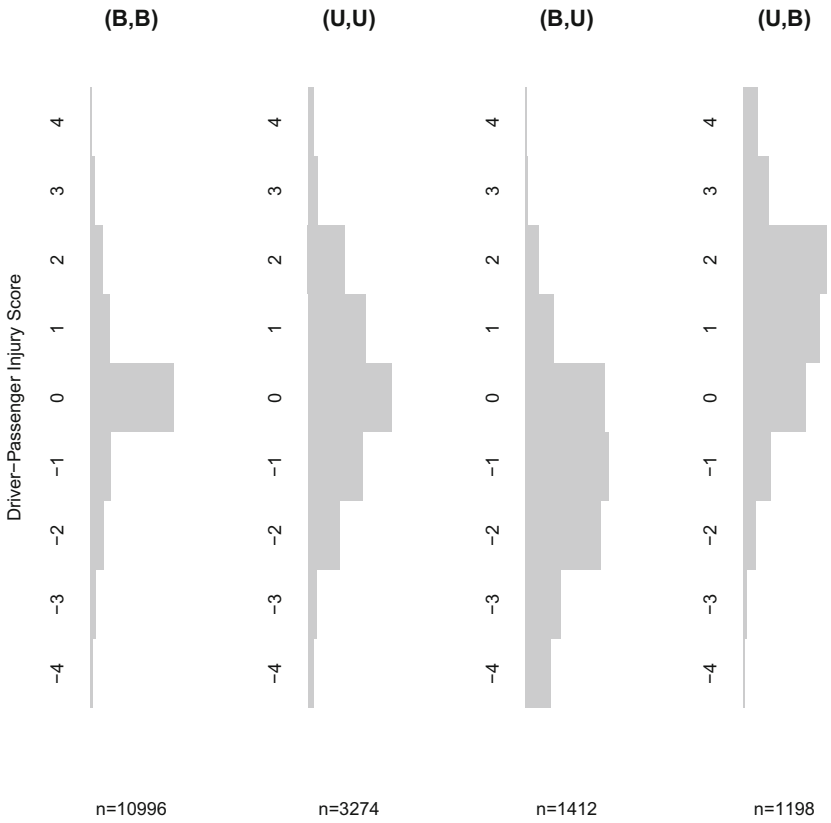

Fig. 7.1 Driver-minus-passenger difference in injury scores for different patterns of safety belt use. The belt use is indicated as (Driver, Passenger), so (B,U) signifies that the driver wore a lap-shoulder belt and the passenger was unbelted

Notably, in Fig. 7.1, to the eye, the distribution of differences in injury scores looks symmetric about zero when driver and passenger are both belted, (B,B). Again, to the eye, when both are unbelted, (U,U), the distribution still looks symmetric about zero, though more dispersed about zero. When the driver is belted and the passenger is not, (B,U), the passenger tended to experience more severe injuries. When the driver is unbelted and the passenger is belted, (U,B), the driver tended to experience more severe injuries. In Fig. 7.1, injury severity seems to track belt use. The eye is an imperfect guide to Fig. 7.1: situations (B,B) and (U,U) exhibit small but statistically significant asymmetries [10, Table 1]. In contrast, the ostensible protective effects of safety belts in situations (B,U) and (U,B) are large and insensitive to large unmeasured biases in who wears safety belts, each with an upper bound ≤ 0.0125 on the one-sided P-value at $\Gamma = 5$; see [10, Table 1].

In brief, before entertaining counterclaims, the initial claim that safety belts reduce the severity of injuries is supported by fairly strong evidence.

7.2.2 The Counterclaim of Selection Bias and Secondary Outcomes

A common counterclaim asserts that the treatment is entirely without effect and the observed associations are the result of self-selection in who receives the treatment. This counterclaim says that if treatment assignments were forcibly reversed in matched pairs, then the only thing that would change is the labeling of individuals as treated or control. How does the situation change if a critic asserts this counterclaim?

In Fig. 7.1, this counterclaim asserts that safety belts have no effect on what happens in a crash: it is simply that frail, injury-prone individuals are disinclined to wear their safety belts. Strap them in and unbelt their companions and nothing will change in the crash, or so the counterclaim asserts. According to this counterclaim: Reverse the belting in the (B,U) or (U,B) crashes and you change the treatment labels, B or U, but nothing about the crash or the injured individuals changes.

As in Sect. 2.3, in the (B,U) and (U,B) situations, write Y_i for the unbelted-minus-belted difference in injury scores.[3] The Y_i tend to be positive in Fig. 7.1, but the counterclaim explains this in terms of selection bias rather than any safety effect of belting. Among other things, the counterclaim asserts that reversing the belting changes only the B or U labels, reversing the sign of Y_i; however, aside from the labels, it does not affect what happens in the crash.

By attributing everything to self-selection in who wears safety belts, this counterclaim denies effects on injuries and on other aspects of the crash. A secondary outcome is the unbelted-minus-belted difference $\widetilde{Y}_i$ in the binary indicators of ejection from the vehicle. Here, $\widetilde{Y}_i = 1$ if only the unbelted individual was ejected, $\widetilde{Y}_i = -1$ if only the belted individual was ejected, and $\widetilde{Y}_i = 0$ if both individuals suffered the same fate. By attributing everything to self-selection, the counterclaim is denying an effect on $\widetilde{Y}_i$ as well, saying that reversing the belting simply changes the sign of $\widetilde{Y}_i$ by changing the B or U labels.

Both Y_i and $\widetilde{Y}_i$ are outcomes, and both the investigator and the critic are interested in injuries, Y_i, not especially in ejections, $\widetilde{Y}_i$. Nonetheless, the counterclaim says that safety belts are without safety effects, and that misleading associations are produced by self-selection in who wears safety belts. It would be a substantial retreat from this counterclaim if the critic admitted that safety belts do cause a reduction in ejections.

The investigator believes that safety belts may have safety effects, perhaps reducing injuries and ejections. When testing the hypothesis of no effect on injuries, Y_i, the investigator views $\widetilde{Y}_i$ as an outcome potentially affected by the treatment. Consistent with these beliefs, the investigator does not adjust for $\widetilde{Y}_i$ when testing no effect on injuries, Y_i, because adjusting for an outcome can introduce bias into a study otherwise free of bias [9].

The counterclaim opens the possibility of an additional analysis. The counterclaim asserts that when treatments are reversed, ejection $\widetilde{Y}_i$ simply changes signs, so

[3]This notation is slightly but not consequentially different from [10].

that its absolute value, $|\widetilde{Y}_i|$, is fixed, unaltered when treatment assignments change. This frees the investigator to conduct an analysis that would be unjustified by the investigator's own beliefs, namely an analysis that acts as if $|\widetilde{Y}_i|$ were a covariate—a fixed attribute of the crash—rather than an outcome.

7.2.3 A Second Look at an Example in Light of a Counterclaim

Figure 7.2 shows one analysis that separates the investigator and the critic. This is an analysis that would be justified were the counterclaim true. The analysis looks at the subset of crashes in which precisely one person was ejected, $|\widetilde{Y}_i| = 1$. Note carefully that the analysis does not specify who was ejected, but merely that one person was ejected and the other was not.

The crashes in Fig. 7.2 are a small subset of the crashes in Fig. 7.1. Nonetheless, the ostensible effect of safety belts on injury severity is larger in Fig. 7.2. In Fig. 7.1, the mean of Y_i is about 1 for both (B,U) and (U,B) pairs, but in Fig. 7.2 the mean increases to more than 1.5. Larger effects are often insensitive to larger unmeasured biases. When testing no effect on injury Y_i in Fig. 7.2, the upper bound on the P-value is ≤ 0.0322 for both (B,U) and (U,B) pairs for $\Gamma' = 11$, substantially higher than $\Gamma = 5$ for Fig. 7.1; see [10, Tables 1 and 2].

The comparison in Fig. 7.2 is insensitive to much larger biases from self-selection than the comparison in Fig. 7.1. And yet, the comparison in Fig. 7.2 is only justified by assuming the counterclaim that associations are the consequence of self-selection. In this sense, the counterclaim undermines itself; it fails in its role as a counterclaim. Like the premise in a proof by contraction, belief in the premise lands the believer in a tangle best avoided.

At this point, the critic might attempt a hasty retreat: "Of course safety belts have causal effects, of course they prevent ejections, but that is entirely without consequence for injuries. Frail, injury-prone individuals decline to wear their safety belts, and so are more often ejected from cars, but it is entirely their frailty, not being hurled from the car into oncoming traffic, that produces an association between belt use and injury severity." Depending upon the context and the nature of the secondary outcome, $\widetilde{Y}_i$, this might be judged either an inconsequential or a major retreat. In either case, the result is a new counterclaim, open to parallel analyses using additional secondary outcomes. See [10, §3.2] for another counterclaim analysis for this example using a different secondary outcome in a different way.

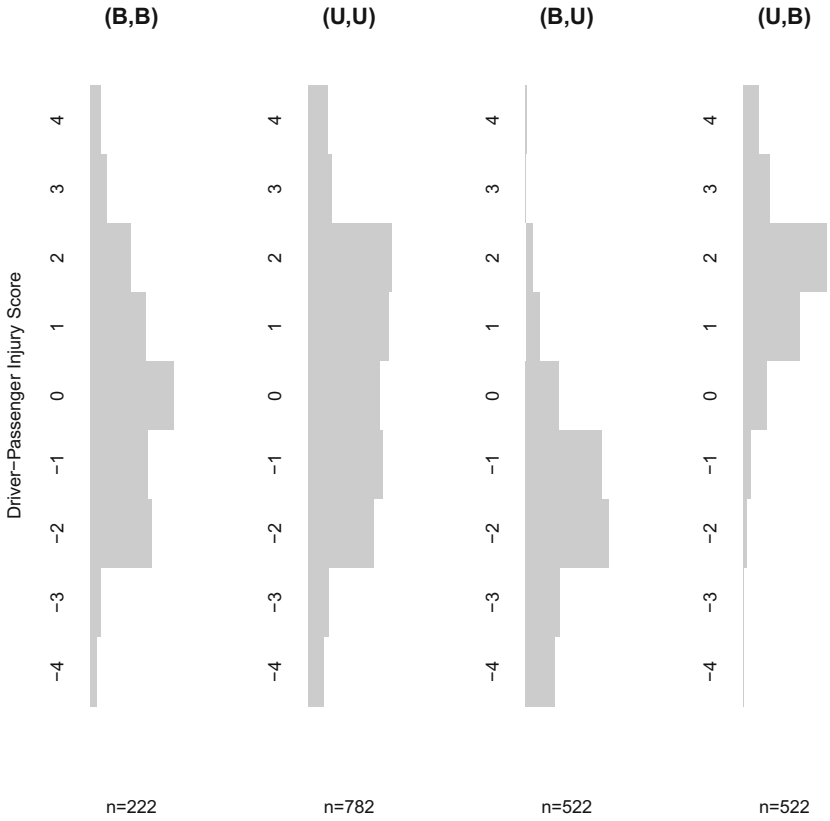

Fig. 7.2 Driver-minus-passenger difference in injury scores for different patterns of safety belt use for crashes in which exactly one of them was ejected from the vehicle. The belt use is indicated as (Driver, Passenger), so (B,U) signifies that the driver wore a lap-shoulder belt and the passenger was unbelted

7.3 Discussion

7.3.1 Anticipating Counterclaims

An investigator anticipates that critics will raise certain counterclaims. In light of this, the investigator conducts analyses to determine whether these anticipated counterclaims undermine themselves. Debate is replaced by anticipated debate. These analyses may demonstrate that certain lines of criticism are unpromising.

7.3.2 *Some Theory*

In Sect. 7.2, the analysis of Fig. 7.2 was insensitive to a larger bias, $\Gamma' = 11$, than the analysis of Fig. 7.2, $\Gamma' = 5$. This pattern may or may not occur—a counterclaim may or may not undermine itself. When is this pattern expected?

The methods in Part III of this book provide an answer in terms of design sensitivity, the limiting sensitivity to unmeasured bias as the sample size increases. Under a simple model for a mediation effect [6], the pattern in Figs. 7.1 and 7.2 is the anticipated pattern: we expect greater insensitivity to bias in large samples in Fig. 7.2; see [10, §4.2]. To emphasize, the validity of the argument in Sect. 7.2 does not depend, in any way, on a mediation model; however, an increase in insensitivity to bias in Fig. 7.2 is anticipated when such a model is true. The sample size in Fig. 7.2 is much smaller than in Fig. 7.1, and this also affects the conclusions of a sensitivity analysis. A simulation shows that the sensitivity analysis of Fig. 7.2 will often have higher power than that of Fig. 7.1, despite the reduced sample size [10, §4.3].

7.4 Further Reading

The main reference for this chapter is [10].

7.5 Data

In the `R` package `DOS2`, the data discussed in this chapter and in [10] are in `frontseat`. The help-file for `frontseat` reproduces several analyses from [10].

Bibliography

1. Austin, J.L.: Philosophical Papers. Oxford, New York (1979)
2. Bross, I.D.J.: Statistical criticism. Cancer **13**, 394–400 (1961). Reprinted with Discussion in Obs. Stud. **4**, 1–70 (2018)
3. Evans, L.: The effectiveness of safety belts in preventing fatalities. Accid. Anal. Prev. **18**, 229–241 (1986)
4. Fisher, R.A.: The effect of methods of ascertainment upon the estimation of frequencies. Ann. Eugen. **6**, 13–25 (1934)
5. Hill, A.B.: The environment and disease: Association or causation? Proc. R. Soc. Med. **58**, 295–300 (1965)
6. Imai, K., Keele, L., Yamamoto, T: Identification, inference and sensitivity analysis for causal mediation effects. Stat. Sci. **25**, 51–71 (2010)

7. Rindskopf, D.: Plausible rival hypotheses in measurement design. In: Bickman, L. (ed.) Research Design: Donald Campbell's Legacy, Vol. 2, pp. 1–12. Sage, Thousand Oaks, CA (2000)
8. Rozelle, R.M., Campbell, D.T.: More plausible rival hypotheses in the cross-lagged panel correlation technique. Psych. Bull. **71**, 74–80 (1969)
9. Rosenbaum, P.R.: The consequences of adjustment for a concomitant variable that has been affected by the treatment. J. R. Stat. Soc. A **147**, 656–666 (1984)
10. Rosenbaum, P.R.: Some counterclaims undermine themselves in observational studies. J. Am. Stat. Assoc. **110**, 1389–1398 (2015)
11. Rosenbaum, P.R., Small, D.S.: Beyond statistical criticism. Obs. Stud. **4**, 63–70 (2018)
12. Weed, D.L., Hursting, S.D.: Biologic plausibility in causal inference: current method and practice. Am. J. Epidemiol. **147**, 415–425 (1998)
13. Weiss, N.: Can the 'specificity' of an association be rehabilitated as a basis for supporting a causal hypothesis? Epidemiol **13**, 6–8 (2002)

Part II
Matching

Chapter 8
A Matched Observational Study

Abstract As a prelude to several chapters describing the construction of a matched control group, the current chapter presents an example of a matched observational study as it might (and did) appear in a scientific journal. When reporting a matched observational study, the matching methods are described very briefly in the *Methods* section. In more detail, the *Results* section presents tables or figures showing that the matching has been effective in balancing certain observed covariates, so that treated and control groups are comparable with respect to these specific variables. The *Results* section then compares outcomes in treated and control groups. Because matching has arranged matters to compare ostensibly comparable groups, the comparison of outcomes is often both simpler in form and more detailed in content than it might be if separate adjustments were required for each aspect of each outcome. Treated and control groups that appear comparable in terms of a specific list of measured covariates—groups that are ostensibly comparable—may nonetheless differ in terms of covariates that were not measured. Though not discussed in the current chapter, the important issue of unmeasured covariates in this example is discussed in Part III.

8.1 Is More Chemotherapy More Effective?

Jeffrey Silber, Dan Polsky, Richard Ross, Orit Even-Shoshan, Sandy Schwartz, Katrina Armstrong, Tom Randall, and I [6, 8] asked how the intensity of chemotherapy for ovarian cancer affected patient outcomes. We thought that greater intensity might prolong survival, perhaps at the cost of increased toxicity. What evidence bears on this question?

There is a basic difficulty in studying the intended effects of medical treatments outside of randomized controlled clinical trials [9]. In virtually all areas of medicine, most of the variation in treatment occurs in thoughtful and deliberate response to variation in the health, prognosis, or wishes of patients. That is, treatment assignment is very far from being determined "at random." Ovarian cancer is unusual in this regard because there is a meaningful source of variation in treatment that is not a response to the patient. Chemotherapy for ovarian cancer is provided by two distinct

© Springer Nature Switzerland AG 2020

P. R. Rosenbaum, *Design of Observational Studies*, Springer Series in Statistics,
https://doi.org/10.1007/978-3-030-46405-9_8

specialties, namely medical oncologists (MOs), who provide chemotherapy for cancers of all kinds, and gynecologic oncologists (GOs), who primarily treat cancers of the ovary, uterus, and cervix. Medical oncologists typically have a residency in internal medicine, followed by a fellowship emphasizing the administration of chemotherapy and the management of its side effects. Gynecologic oncologists typically complete a residency in obstetrics and gynecology, followed by a fellowship in gynecologic oncology, that includes training in surgical oncology and chemotherapy administration for gynecologic cancers. Unlike gynecologic oncologists, who are trained in surgery, medical oncologists are almost invariably not surgeons, so medical oncologists provide chemotherapy after someone else has performed surgery. We anticipated, as turned out to be the case, that MOs would use chemotherapy more intensively than GOs, both at the time of initial diagnosis and several years later if the cancer has spread from its site of origin. Is the greater intensity of chemotherapy found in the practice of MOs of benefit to patients?

The study was based on data that linked the Surveillance, Epidemiology and End Results (SEER) program of the U.S. National Cancer Institute to Medicare claims. The SEER data are collected at SEER sites in the USA, where some SEER sites are cities (e.g., Detroit) and others are states (e.g., New Mexico). The SEER data include clinical stage and tumor grade. The study used data on women older than 65 with ovarian cancer who were diagnosed between 1991 and 1999 and had appropriate surgery and at least some chemotherapy; see [8] for details. There were 344 such women who received chemotherapy from a GO, and 2011 such women who received chemotherapy from an MO.

8.2 Matching for Observed Covariates

We matched each GO patient to an MO patient, producing 344 matched pairs of two similar patients. Table 8.1 describes several of the covariates used in the matching. Was the matching successful in producing GO and MO groups that were reasonably comparable in terms of observed covariates?

The first group of variables in Table 8.1 describes the type of surgeon. Gynecologic oncologists are specialists in surgery for gynecologic cancers, but surgery is often performed by gynecologists or general surgeons. Not surprisingly, this variable is substantially out of balance before matching: a patient who has surgery performed by a GO is more likely to have chemotherapy provided by a GO. After matching, the distributions of surgeon types are almost identical.

Clinical stage is usually thought to be the most important predictor of survival in ovarian cancer. Perhaps because there is no particular reason for patients to seek chemotherapy from an MO or a GO, stage was not greatly out of balance before matching. For stage, the balance after matching is quite close.

In fact, whenever possible, we matched exactly for surgeon type and stage. This means that, whenever possible, a stage III patient whose surgery was performed by a gynecologist was paired with a stage III patient whose surgery was performed

Table 8.1 Comparability at baseline for patients with ovarian cancer: all 344 patients of a gynecologic oncologist (GO), 344 matched patients of a medical oncologist (MO), and all 2011 patients of medical oncologists. Values are in percent, unless labeled as means. Also included in the matching, but not in this table, were additional comorbid conditions, including anemia, angina, arrhythmia, asthma, coagulation disorder, electrolyte abnormality, hepatic dysfunction, hyperthyroidism, peripheral vascular disease, and rheumatoid arthritis. Notice that surgeon type, SEER site, and year of diagnosis were substantially out of balance prior to matching, but were in reasonable balance after matching

		GO	matched-MO	all-MO
		$n = 344$	$n = 344$	$n = 2,011$
Surgeon	GO	76	75	33
Type	Gyn	15	16	39
	General	8	8	28
Stage	I	9	9	9
	II	11	9	9
	III	51	53	47
	IV	26	26	31
	Missing	3	2	3
Tumor	1	5	4	4
Grade	2	16	13	17
	3	52	55	47
	4	9	8	11
	Missing	18	20	21
Demographics	Age, mean	72.2	72.2	72.8
	White	91	94	94
	Black	8	5	3
Selected	COPD	15	12	13
Comorbid	Hypertension	48	46	42
Conditions	Diabetes	11	8	8
	CHF	2	2	4
SEER Site	Connecticut	18	18	15
	Detroit	26	26	12
	Iowa	17	17	17
	New Mexico	7	7	3
	Seattle	9	9	16
	Atlanta	9	9	7
	Los Angeles	12	12	19
	San Francisco	1	1	9
Year of Diagnosis	1991	4	4	9
	1992	7	7	14
	1993	10	9	14
	1994	11	11	12
	1995	11	13	12
	1996	10	9	12
	1997	16	15	10
	1998	13	15	9
	1999	18	17	9
Propensity Score	$\widehat{e}(\mathbf{x})$, mean	0.23	0.21	0.14

by a gynecologist. As seen in the marginal distributions in Table 8.1, this was not always possible. Naive intuition suggests that one should match exactly for every variable, but a moment's thought reveals this to be impossible. If there were just 30 binary variables, there would be 2^{30} or about a billion types of patients, so it is highly unlikely that one would find exact matches on all 30 variables for almost all of the patients. Nonetheless, the balance on the covariates in Table 8.1 is quite good, even though, unlike stage and surgeon type, most variables in Table 8.1 are not exactly matched. Section 10.1 will discuss covariate balance in greater detail. In practice, exact matching, if used at all, is reserved for one or two critically important variables.

Surgeon type and clinical stage are connected. A surgeon who is trained specifically in cancer surgery, such as a GO, is likely to combine an effort at curative surgery with extensive node sampling. In consequence, a cancer surgeon may find more cancer than a general surgeon, and may assign a patient to a higher clinical stage [2]. That is, a patient who is classified as stage II by a general surgeon might have been classified as stage III by a GO. The very meaning of one variable, clinical stage, depends upon another variable, surgeon type. However, because almost all of the pairs are matched exactly for stage and surgeon type, in almost all pairs, the two patients received the same stage from the same category of surgeon.

Tumor grade, demographics, and selected comorbid conditions are seen to be reasonably well balanced, and the same is true for other comorbid conditions mentioned in the caption of Table 8.1. Comorbid conditions matter both because they may pose a direct risk to the patient's survival and may also complicate and limit the administration of chemotherapy.

Gynecologic oncologists are unevenly spread through the USA, and the profession has been growing over time. For this reason, before matching, both SEER site and year of diagnosis were often quite different for the patients of GOs and MOs. For patients of GOs at SEER sites, 26% were in Detroit and only 1% in San Francisco, while for all MO patients, 12% were in Detroit and 9% in San Francisco. Expressed in terms of odds ratios, a woman is almost 20 times as likely to be treated by a GO in Detroit than in San Francisco. Although ovarian cancer is, presumably, much the same disease in Detroit and San Francisco, there are differences in wealth, demographics, and health services that might affect patient outcomes. After matching, the distributions of SEER sites are the same. Similarly, GO patients were more likely to have been diagnosed recently, simply because there are more GOs recently. Although ovarian cancer itself may not have changed much through the 1990s, the chemotherapies did improve [3], so a substantial imbalance in year of diagnosis would compare patients who received different drugs. After matching, the proportion of patients diagnosed in three time intervals, 1991–1992, 1993–1996, and 1997–1999, were identical, with only small imbalances for individual years inside these intervals. In fact, although this is not visible in Table 8.1, the joint behavior of the three time intervals and the SEER site is exactly balanced; for instance, the number of GO patients in 1991–1992 from Iowa equals the number of matched MO patients in 1991–1992 from Iowa. The balance on SEER site and year of diagnosis was obtained using "fine balance" [5, 6], a technique discussed in Chap. 11.

The large imbalances before matching in surgeon type, year of diagnosis, and SEER site, and the much smaller imbalances before matching in clinical stage, grade and comorbid conditions are somewhat encouraging. They hint at the possibility that the chemotherapy provider type reflects the relative availability of MOs and GOs more than it reflects attributes of the patient's disease. In any event, the measured covariates are reasonably well balanced after matching.

The final variable in Table 8.1 is the propensity score. It is an estimate of the conditional probability of receiving chemotherapy from a GO rather than an MO given the observed covariates. Propensity scores are a basic tool when matching on many variables [4]. Propensity scores were discussed conceptually in Sect. 3.3, and their role in multivariate matching is discussed further in Chap. 9. In Table 8.1, the mean propensity score is fairly similar in the two matched groups; indeed, the balance on many covariates is produced by balancing the propensity score; see (3.10).

8.3 Outcomes in Matched Pairs

As anticipated, many MOs often used chemotherapy more intensely than did GOs, both in the first year following diagnosis and in the first five years following diagnosis; see Table 8.2 and Fig. 8.1. The difference at the medians is not large, but it is quite noticeable at the upper quartiles. Was greater intensity of treatment by MOs beneficial?

Despite a difference in intensity, survival was virtually identical for the patients of MOs and GOs; see Table 8.3. The standard test comparing paired censored survival times is the Prentice-Wilcoxon test [7]; it gives a two-sided P-value of 0.45. Table 8.3 also lists the number of patients at risk at the beginning of each of the first five years, that is, the number of patients alive and uncensored at the start of the year. The number of patients at risk is also very similar for matched patients of MOs and GOs.

Patients of MOs had more weeks of toxicity than did patients of GOs; see Table 8.4 and Fig. 8.1. In Table 8.4, both in year 1 and in years 1 to 5, the upper quartiles for MO patients are double those for GO patients.

Table 8.2 Weeks with chemotherapy, in year 1 and in years 1 through 5, for 344 matched pairs. Values are the mean, quartiles, minimum, and maximum. Two-sided P-values are from Wilcoxon's signed rank test

Period	Group	Mean	Min	25%	50%	75%	Max	P-value
Year 1	GO	6.63	1	5	6	8	19	0.0022
Year 1	MO	7.74	1	5	6	10	42	
Years 1–5	GO	12.07	1	5	9	16	70	0.00045
Years 1–5	MO	16.47	1	6	11	21	103	

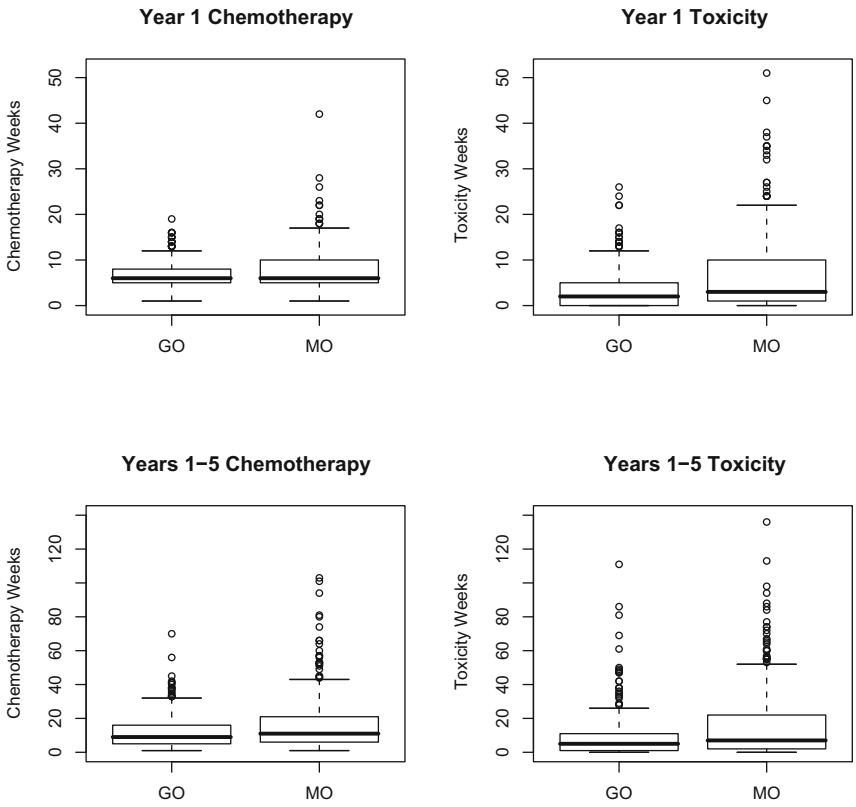

Fig. 8.1 Chemotherapy and toxicity in 344 pairs of patients with ovarian cancer, one treated by a GO, the other by an MO

The difference in intensity was associated with the difference in toxicity; see Fig. 8.2. In Fig. 8.2, the 344 matched pair differences in toxicity weeks, MO-minus-GO, are plotted against the 344 matched pair differences in chemotherapy weeks. Points project horizontally into the marginal boxplot of the differences in toxicity weeks. The curve is a lowess smooth [1, pages 168–180] as implemented in the statistical package R. Kendall's rank correlation between toxicity differences and chemotherapy differences is 0.39 and is significantly different from zero, with a two-sided significance level less than 10^{-10}.

In brief, it appears that MOs often treated more intensively than GOs, often with more toxicity, but survival was no different.

An obvious question is whether the patterns exhibited in Tables 8.2, 8.3, 8.4 and Figs. 8.1 and 8.2 could be the result not of differences in intensity of chemotherapy applied to similar patients, but rather of unmeasured pretreatment differences between the patients themselves. Table 8.1 showed that, prior to matching, many covariates were substantially out of balance in the GO and MO groups. It is certainly

Table 8.3 Survival and number at risk in 344 matched pairs of one GO patient and one MO patient. The two-sided P-value of 0.45 is from the Prentice-Wilcoxon test [7] comparing paired survival times

	GO Patients	MO Patients
Median (Years)	3.04	2.98
95% CI	[2.50, 3.40]	[2.69, 3.67]
1 Year Survival %	86.6	87.5
95% CI	[83.0, 90.2]	[84.0, 90.1]
2 Year Survival %	64.8	66.9
95% CI	[59.8, 69.9]	[61.9, 71.8]
5 Year Survival %	35.1	34.2
95% CI	[30.0, 40.2]	[29.2, 39.3]
Number at Risk Year 0	344	344
Number at Risk Year 1	298	301
Number at Risk Year 2	223	230
Number at Risk Year 3	173	172
Number at Risk Year 4	133	128

Table 8.4 Weeks with chemotherapy-associated toxicity, in year 1 and in years 1 through 5, for 344 matched pairs. Values are the mean, quartiles, minimum, and maximum. Two-sided P-values are from Wilcoxon's signed rank test. Chemotherapy-associated toxicity refers to inpatient or outpatient diagnoses for anemia, neutropenia, thrombocytopenia, diarrhea, dehydration or mucositis, and neuropathy

Period	Group	Mean	Min	25%	50%	75%	Max	P-value
Year 1	GO	3.61	0	0	2	5	26	0.00000089
Year 1	MO	6.67	0	1	3	10	51	
Years 1–5	GO	8.89	0	1	5	11	111	0.000000026
Years 1–5	MO	16.29	0	2	7	22	136	

possible that some other covariate, a covariate that was not measured, was also out of balance before matching, and because that variable was not controlled by matching, it is certainly possible that the variable is still out of balance in Tables 8.2, 8.3, 8.4 and Figs. 8.1 and 8.2. This concern is central to Part III, where there will be further analysis of the ovarian data.

8.4 Summary

This chapter is an introduction to Part II. A matched comparison has been presented as it might be, and was, presented in a scientific journal [8], with little reference to the procedures used to construct the matched sample [6]. Matching procedures are the focus of Part II. The current chapter has made three basic points. First, matching on many observed covariates is often feasible; see Table 8.1. Second,

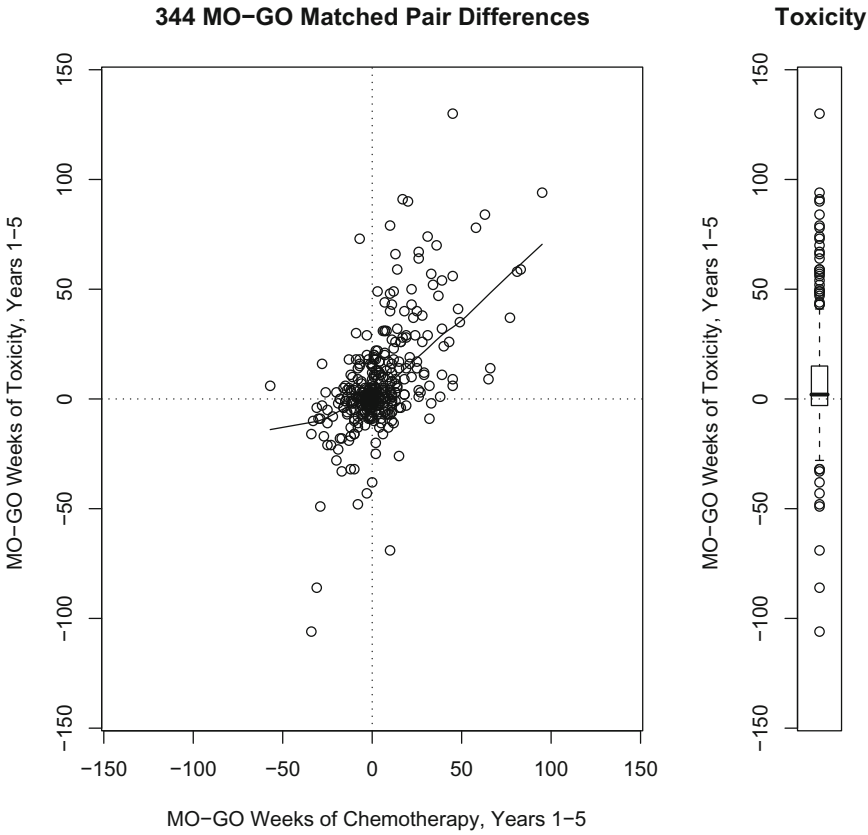

Fig. 8.2 Matched pair differences, MO−GO, in toxicity plotted against differences in chemotherapy. The curve is a lowess smooth

the reader may examine the degree to which matched groups are comparable with respect to observed covariates, as well as which covariates are not among the observed covariates, without getting involved in the procedures used to construct the matched sample; again, see Table 8.1. Finally, straightforward analyses sufficed to take a close look at several outcomes—here, survival, chemotherapy, toxicity in Tables 8.2, 8.3, 8.4 and Figs. 8.1, 8.2—because those analyses compared two matched groups that looked similar prior to treatment with respect to observed covariates. More precisely, analyses conducted under the naïve model for treatment assignment (Sect. 3.3) are straightforward. The possibility that the naïve model is not true is taken up in Part III.

8.5 Further Reading

The remainder of Part II is further reading for this chapter. The ovarian cancer study is discussed in detail by Jeffrey Silber and colleagues [6, 8].

References

1. Cleveland, W.S.: The Elements of Graphing Data. Hobart Press, Summit, NJ (1994)
2. Mayer, A.R., Chambers, S.K., Graves, E., et al.: Ovarian cancer staging: Does it require a gynecologic oncologist? Gynecol. Oncol. **47**, 223–337 (1992)
3. McGuire, W.P., Bundy, B., Wenzel, L., et al.: Cyclophosphamide and cisplatin compared with paclitaxel and cisplatin in patients with stage III and stage IV ovarian cancer. N. Engl. J. Med. **334**, 1–6 (1996)
4. Rosenbaum, P.R., Rubin, D.B.: The central role of the propensity score in observational studies for causal effects. Biometrika **70**, 41–55 (1983)
5. Rosenbaum, P.R.: Optimal matching in observational studies. J. Am. Stat. Assoc. **84**, 1024–1032 (1989)
6. Rosenbaum, P.R., Ross R.N., Silber, J.H.: Minimum distance matched sampling with fine balance in an observational study of treatment for ovarian cancer. J. Am. Stat. Assoc. **102**, 75–83. (2007)
7. O'Brien, P.C., Fleming, T.R.: A paired Prentice-Wilcoxon test for censored paired data. Biometrics **43**, 169–180 (1987)
8. Silber, J.H., Rosenbaum, P.R., Polsky, D., Ross, R.N., Even-Shoshan, O., Schwartz, S., Armstrong, K.A., Randall, T.C.: Does ovarian cancer treatment and survival differ by the specialty providing chemotherapy? J. Clin. Oncol. **25**, 1169–1175 (2007); Related editorial **25**, 1157–1159 (2007); Related letters and rejoinders **25**, 3552–3558 (2007)
9. Vandenbroucke, J.P.: When are observational studies as credible as randomized trials? Lancet **363**, 1728–1731 (2004)

Chapter 9
Basic Tools of Multivariate Matching

Abstract The basic tools of multivariate matching are introduced, including the propensity score, distance matrices, calipers imposed using a penalty function, optimal matching, matching with multiple controls, and full matching. The tools are illustrated with a tiny example from genetic toxicology ($n = 46$), an example that is so small that one can keep track of individuals as they are matched using different techniques.

9.1 A Small Example

The mechanics of matching are best illustrated with a very small example, not because such an example is representative but because it is possible to inspect the details of what goes on. The example considered here has 47 subjects and three covariates. Typical examples have many more subjects and many more covariates.

Welders are exposed to chromium and nickel, substances that can cause inappropriate links between DNA and proteins, which in turn may disrupt gene expression or interfere with replication of DNA. Costa, Zhitkovich, and Toniolo [14] measured DNA–protein cross-links in samples of white blood cells from 21 railroad arc welders exposed to chromium and nickel and 26 unexposed controls. All 47 subjects were male. In their data in Table 9.1, there are three covariates, namely age, race, and current smoking behavior. The response is a measure of DNA–protein cross-links.

Summary measures describing the covariates appear at the bottom of Table 9.1. The welders are about five years younger than the controls on average, have relatively fewer African Americans, and more smokers. For what it is worth, the difference in age between welders and controls is significant as judged by the t-test, with a t of -2.25 and a two-sided significance level of 0.03. The two binary variables yield t's of -0.94 for race and 1.2 for smoking, and are neither significant by that standard nor by Fisher's exact test for a 2×2 table. In an experiment, random assignment leads to the expectation that one covariate in 20 will exhibit

© Springer Nature Switzerland AG 2020

P. R. Rosenbaum, *Design of Observational Studies*, Springer Series in Statistics,
https://doi.org/10.1007/978-3-030-46405-9_9

Table 9.1 Unmatched data for 21 railroad arc welders and 26 potential controls. Covariates are age, race (C = Caucasian , AA = African American), current smoker (Y = yes, N = no). Response is DPC = DNA–protein cross-links in percent in white blood cells. All 47 subjects are male

Welders					Controls				
ID	Age	Race	Smoker	DPC	ID	Age	Race	Smoker	DPC
1	38	C	N	1.77	1	48	AA	N	1.08
2	44	C	N	1.02	2	63	C	N	1.09
3	39	C	Y	1.44	3	44	C	Y	1.10
4	33	AA	Y	0.65	4	40	C	N	1.10
5	35	C	Y	2.08	5	50	C	N	0.93
6	39	C	Y	0.61	6	52	C	N	1.11
7	27	C	N	2.86	7	56	C	N	0.98
8	43	C	Y	4.19	8	47	C	N	2.20
9	39	C	Y	4.88	9	38	C	N	0.88
10	43	AA	N	1.08	10	34	C	N	1.55
11	41	C	Y	2.03	11	42	C	N	0.55
12	36	C	N	2.81	12	36	C	Y	1.04
13	35	C	N	0.94	13	41	C	N	1.66
14	37	C	N	1.43	14	41	AA	Y	1.49
15	39	C	Y	1.25	15	31	AA	Y	1.36
16	34	C	N	2.97	16	56	AA	Y	1.02
17	35	C	Y	1.01	17	51	AA	N	0.99
18	53	C	N	2.07	18	36	C	Y	0.65
19	38	C	Y	1.15	19	44	C	N	0.42
20	37	C	N	1.07	20	35	C	N	2.33
21	38	C	Y	1.63	21	34	C	Y	0.97
					22	39	C	Y	0.62
					23	45	C	N	1.02
					24	42	C	N	1.78
					25	30	C	N	0.95
					26	35	C	Y	1.59
	Mean	AA	Smoker			Mean	AA	Smoker	
	Age	%	%			Age	%	%	
	38	10	52			43	19	35	

an imbalance that is statistically significant at the 0.05 level, at least when the comparison is based on an appropriate randomization test, such as Fisher's exact test.[1]

[1]Cochran [12] discusses the relationship between the magnitude of a t-statistic for covariate imbalance and the coverage rate of a confidence interval for a treatment effect when no adjustment is made for the covariate. He concludes that problems begin to occur before the conventional 0.05 level of significance is reached, and that attention should be given to covariates exhibiting a t-statistic of 1.5 in absolute value.

A pair matching would form 21 pairs of a welder and a control, discarding five potential controls, in the process altering the marginal distribution of the three covariates in the control group. Pair matching is, perhaps, not the most attractive method in this particular example. In part, the removal of five controls can have only a moderate impact on the distribution of covariates. In part, in such a small example, it is not entirely comfortable to discard five controls. Then again, the oldest control, #2, is 63 years old, which is ten years older than the oldest welder, #18, and at least 19 years older than all other welders; arguably, some controls are not suitable for comparison.

As Rubin [47] observes, there is a definite limit to what pair matching can accomplish. In many cases, what is needed falls well within this attainable limit; in other cases it does not. Importantly, to a large extent, this limit is not shared by more flexible matching strategies, such as "full matching" [41]. One can see the limits of pair matching in Table 9.2. The mean age of the 21 welders is 38.2 years, while the mean age of the 26 controls is 42.7 years. Even if the only consideration were reducing the mean age in the control group, in this example, pair matching will fall slightly short of the goal, because the mean age of the 21 youngest of the 26 controls is 39.6 years. In parallel, for smoking, 11 of the 21 welders are smokers, while only 9 of the 26 potential controls are smokers, so pair matching can at most select all nine smoking controls, and cannot perfectly balance smoking. Moreover, the 21 youngest controls include only eight of the nine smokers. Furthermore, although pair matching can eliminate the imbalance in race, if it does this, it cannot use all nine smoking controls. The discrepancies are not large, and hence are perhaps tolerable in this example, but in other studies, a matching strategy more flexible than pair matching may be needed.

Before matching, there are $L = 47$ subjects in Table 9.1, $\ell = 1, 2, \ldots, L$. In the example, for subject ℓ, the observed covariate, $\mathbf{x}_\ell$, is three-dimensional, $\mathbf{x}_\ell = (x_{\ell 1}, x_{\ell 2}, x_{\ell 3})^T$, where (i) $x_{\ell 1}$ is the age of subject ℓ, (ii) $x_{\ell 2}$ encodes race, $x_{\ell 2} = 1$ if ℓ is African American, $x_{\ell 2} = 0$ if ℓ is Caucasian, (iii) $x_{\ell 3}$ encodes smoking, $x_{\ell 3} = 1$ if ℓ is a current smoker, $x_{\ell 3} = 0$ otherwise. For instance, $\mathbf{x}_1 = (38, 0, 0)^T$. The variable Z_ℓ distinguishes treated subjects from potential controls: $Z_\ell = 1$ for a treated subject, here a welder; $Z_\ell = 0$ for a potential control.

Table 9.2 Limitations of pair matching for age in the welder data. Even the 21 youngest of the 26 controls are, on average, slightly older than the welders, so even a pair matching that focused exclusively on one covariate, age, could not entirely eliminate the age difference. The discrepancy is not large in this example but can be large in some studies. Other forms of matching besides pair matching can make further progress

	Mean	Min	Q1	Median	Q3	Max
All 26 controls	42.7	30	36	42	48	63
21 youngest controls	39.6	30	35	40	44	50
21 welders	38.2	27	35	38	39	53

9.2 Propensity Score

As discussed in Sect. 3.3, the propensity score is the conditional probability of exposure to treatment given the observed covariates, $e(\mathbf{x}) = \Pr(Z = 1 \mid \mathbf{x})$. The propensity score is defined in terms of the observed covariates, $\mathbf{x}$, even though there is invariably concern about other covariates that were not measured. Properties of the propensity score are discussed in Sect. 3.3 and [36, 37, 39, 42, 46].

In the simplest randomized experiment, with treatments assigned by independent flips of a fair coin, the propensity score is $e(\mathbf{x}) = \Pr(Z = 1 \mid \mathbf{x}) = \frac{1}{2}$ for all values of the covariate $\mathbf{x}$. In this case, the covariate $\mathbf{x}$ is not useful in predicting the treatment that a subject will receive. In such a completely randomized experiment, smokers are just as likely as nonsmokers to receive the treatment — the chance is $\frac{1}{2}$ for smokers and $\frac{1}{2}$ for nonsmokers — so smokers tend to show up in the treated group just about as often as they do in the control group, and any difference in smoking frequency is due to chance, the flip of a coin that assigned one subject to treatment, another to control.

Brief examination of Table 9.1 suggested that at least age, x_1, and possibly race and smoking, x_2 and x_3, could be used to predict treatment assignment Z, so that the propensity score is not constant. Saying that welders tend to be somewhat younger than controls is much the same as saying that the chance of being a welder is lower for someone who is older, that is, $e(\mathbf{x})$ is lower when x_1 is higher.

If two subjects have the same propensity score, $e(\mathbf{x})$, they may have different values of $\mathbf{x}$. For instance, a younger nonsmoker and an older smoker might have the same propensity score, $e(\mathbf{x})$, because welders are often younger smokers. Imagine that we have two individuals, a welder and a control, with the same propensity score $e(\mathbf{x})$ but different covariates $\mathbf{x}$. Although these two individuals differ in terms of $\mathbf{x}$, this difference in $\mathbf{x}$ will not be helpful in guessing which one is the welder, because $e(\mathbf{x})$ is the same. If subjects were matched for $e(\mathbf{x})$, they may be mismatched for $\mathbf{x}$, but the mismatches in $\mathbf{x}$ will be due to chance and will tend to balance, particularly in large samples. If young nonsmokers and old smokers have the same propensity score, then a match on the propensity score may pair a young nonsmoking welder to an old smoking control, but it will do this about as often as it pairs an old smoking welder to a young nonsmoking control.

In brief, matching on $e(\mathbf{x})$ tends to balance $\mathbf{x}$; see (3.10) in Sect. 3.3 or [36]. More precisely, treatment assignment Z is conditionally independent of observed covariates $\mathbf{x}$ given the propensity score $e(\mathbf{x})$. Moreover, the propensity score is the coarsest function of $\mathbf{x}$ with this balancing property, so if the propensity score $e(\mathbf{x})$ is not balanced, then the observed covariates $\mathbf{x}$ will continue to be useful in predicting treatment Z. Expressed differently, ignoring chance imbalances, as we might in sufficiently large samples, balancing the propensity score $e(\mathbf{x})$ is sufficient to balance the observed covariates $\mathbf{x}$ but also necessary to balance $\mathbf{x}$. It is important to understand that this is a true statement about observed covariates $\mathbf{x}$, and only about observed covariates, whether or not treatment assignment also depends on covariates that were not measured. On the negative side, success in balancing

the observed covariates, $\mathbf{x}$, provides no assurance that unmeasured covariates are balanced. On the positive side, balancing observed covariates — comparing groups that, at least, look comparable in terms of the observed $\mathbf{x}$ — is a discrete task that can be successfully completed without reference to possible imbalances in covariates that were not measured. For the men in Table 9.1, imagine that having a father who worked as a welder is associated with a much higher chance that the son also will work as a welder, so father's occupation is an unmeasured covariate that could be used to predict treatment Z; then success in matching on the propensity score $e(\mathbf{x})$ would tend to balance age, race, and smoking, but there is no reason to expect it to balance father's occupation. In short, (i) matching on $e(\mathbf{x})$ is often practical even when there are many covariates in $\mathbf{x}$ because $e(\mathbf{x})$ is a single variable, (ii) matching on $e(\mathbf{x})$ tends to balance all of $\mathbf{x}$, and (iii) failure to balance $e(\mathbf{x})$ implies that $\mathbf{x}$ is not balanced.

The propensity score is unknown, but it can be estimated from the data at hand. In the example, the propensity score is estimated by a linear logit model

$$\log\left\{\frac{e(\mathbf{x}_\ell)}{1-e(\mathbf{x}_\ell)}\right\} = \zeta_0 + \zeta_1 x_{\ell 1} + \zeta_2 x_{\ell 2} + \zeta_3 x_{\ell 3}, \tag{9.1}$$

and the fitted values $\widehat{e}(\mathbf{x}_\ell)$ from this model are the estimates of the propensity score. The estimates of the propensity score, $\widehat{e}(\mathbf{x}_\ell)$, are displayed in Table 9.3. Control #2, a 63-year-old Caucasian nonsmoker has $\widehat{e}(\mathbf{x}_\ell) = 0.09$, only a 9% estimated chance of being a welder. In contrast, control #12, a 36-year-old smoker, has $\widehat{e}(\mathbf{x}_\ell) = 0.64$, a 64% chance of being a welder, so this control actually has covariates that are atypical of controls and more typical of welders. Welders #10 and #18 have similar estimated propensity scores $\widehat{e}(\mathbf{x}_\ell)$ but different patterns of covariates $\mathbf{x}_\ell$.

The limitations of pair matching, illustrated for age in Sect. 9.1, apply with any variable, including the propensity score. The mean of the 21 largest $\widehat{e}(\mathbf{x})$'s in the control group is 0.46, somewhat less than the mean of $\widehat{e}(\mathbf{x})$ in the treated group, namely 0.51, so no pair matching can completely close that gap.

9.3 Distance Matrices

In its simplest form, a distance matrix is a table with one row for each treated subject and one column for each potential control. For the welder data in Table 9.3, the distance matrix would have 21 rows and 26 columns; it would be 21×26. The value in row i and column j of the table is the "distance" between the ith treated subject and the jth potential control. This "distance" is a nonnegative number[2] or infinity, ∞, which measures the similarity of two individuals in terms of their covariates

[2]The distance need not be, and typically is not, a distance in the sense that the word is used in metric space topology: it need not satisfy the triangle inequality.

Table 9.3 Estimated propensity scores $\widehat{e}(\mathbf{x})$ for 21 railroad arc welders and 26 potential controls. Covariates are age, race (C = Caucasian, AA = African American), current smoker (Y = yes, N = no)

Welders					Controls				
ID	Age	Race	Smoker	$\widehat{e}(\mathbf{x})$	ID	Age	Race	Smoker	$\widehat{e}(\mathbf{x})$
1	38	C	N	0.46	1	48	AA	N	0.14
2	44	C	N	0.34	2	63	C	N	0.09
3	39	C	Y	0.57	3	44	C	Y	0.47
4	33	AA	Y	0.51	4	40	C	N	0.42
5	35	C	Y	0.65	5	50	C	N	0.23
6	39	C	Y	0.57	6	52	C	N	0.20
7	27	C	N	0.68	7	56	C	N	0.15
8	43	C	Y	0.49	8	47	C	N	0.28
9	39	C	Y	0.57	9	38	C	N	0.46
10	43	AA	N	0.20	10	34	C	N	0.54
11	41	C	Y	0.53	11	42	C	N	0.38
12	36	C	N	0.50	12	36	C	Y	0.64
13	35	C	N	0.52	13	41	C	N	0.40
14	37	C	N	0.48	14	41	AA	Y	0.35
15	39	C	Y	0.57	15	31	AA	Y	0.55
16	34	C	N	0.54	16	56	AA	Y	0.13
17	35	C	Y	0.65	17	51	AA	N	0.12
18	53	C	N	0.19	18	36	C	Y	0.64
19	38	C	Y	0.60	19	44	C	N	0.34
20	37	C	N	0.48	20	35	C	N	0.52
21	38	C	Y	0.60	21	34	C	Y	0.67
					22	39	C	Y	0.57
					23	45	C	N	0.32
					24	42	C	N	0.38
					25	30	C	N	0.63
					26	35	C	Y	0.65
	Mean Age	AA %	Smoker %	Mean $\widehat{e}(\mathbf{x})$		Mean Age	AA %	Smoker %	Mean $\widehat{e}(\mathbf{x})$
	38	10	52	0.51		43	19	35	0.39

$\mathbf{x}$. Two individuals with the same value of $\mathbf{x}$ would have distance zero. An infinite distance in row i and column j is used to forbid matching the ith treated subject to the jth potential control.

For the welder data in Table 9.3, Table 9.4 displays all 21 rows and the first 6 of 26 columns on one such distance matrix. The distance is the squared difference in the estimated propensity score, $\widehat{e}(\mathbf{x})$. The first welder has $\widehat{e}(\mathbf{x}) = 0.46$ and the first control has $\widehat{e}(\mathbf{x}) = 0.14$, so the distance in the first row and first column of Table 9.4 is $(0.46 - 0.14)^2 = 0.10$. If the only concern were to obtain close matches on $\widehat{e}(\mathbf{x})$, then this might be a reasonable distance. The disadvantage is that two controls

Table 9.4 Squared differences in propensity scores between welders and controls. Rows are the 21 welders and columns are for the first 6 of 26 potential controls

Welder	Control 1	Control 2	Control 3	Control 4	Control 5	Control 6
1	0.10	0.13	0.00	0.00	0.05	0.06
2	0.04	0.06	0.02	0.01	0.01	0.02
3	0.19	0.23	0.01	0.02	0.12	0.14
4	0.13	0.18	0.00	0.01	0.08	0.09
5	0.26	0.32	0.03	0.06	0.18	0.20
6	0.19	0.23	0.01	0.02	0.12	0.14
7	0.29	0.35	0.05	0.07	0.20	0.23
8	0.12	0.16	0.00	0.01	0.07	0.08
9	0.19	0.23	0.01	0.02	0.12	0.14
10	0.00	0.01	0.07	0.05	0.00	0.00
11	0.15	0.19	0.00	0.01	0.09	0.11
12	0.13	0.17	0.00	0.01	0.07	0.09
13	0.14	0.19	0.00	0.01	0.08	0.10
14	0.11	0.15	0.00	0.00	0.06	0.08
15	0.19	0.23	0.01	0.02	0.12	0.14
16	0.16	0.20	0.01	0.02	0.10	0.11
17	0.26	0.32	0.03	0.06	0.18	0.20
18	0.00	0.01	0.08	0.05	0.00	0.00
19	0.20	0.25	0.02	0.03	0.13	0.15
20	0.11	0.15	0.00	0.00	0.06	0.08
21	0.20	0.25	0.02	0.03	0.13	0.15

with the same propensity score, $\widehat{e}(\mathbf{x})$, may have different patterns of covariates, $\mathbf{x}$, and this is ignored in Table 9.4. For example, in the first row and third and fourth columns of Table 9.4, the distance is zero to two decimal places. Looking back at Table 9.3, the distances between welder #1 and potential controls #3 and #4 are, respectively, $(0.46 - 0.47)^2 = 0.0001$ and $(0.46 - 0.42)^2 = 0.0016$, so control #3 is ever so slightly closer to welder #1. However, in terms of the details of $\mathbf{x}$, control #4 looks to be the better match, a nonsmoker with a two-year difference in age, as opposed to control #3, a smoker with a six-year difference in age. Because younger smokers are more common in the welder group, the propensity score is indifferent between a younger nonsmoker and an older smoker, but the details of $\mathbf{x}$ suggest that control #4 is a better match for welder #1 than is control #4.

An alternative distance [38] insists that individuals be close on the propensity score, $\widehat{e}(\mathbf{x})$, but once this is achieved, the details of $\mathbf{x}$ affect the distance. With a caliper of width w, if two individuals, say k and ℓ, have propensity scores that differ by more than w — that is, if $|\widehat{e}(\mathbf{x}_k) - \widehat{e}(\mathbf{x}_\ell)| > w$ — then the distance between these individuals is set to ∞, whereas, if $|\widehat{e}(\mathbf{x}_k) - \widehat{e}(\mathbf{x}_\ell)| \leq w$, the distance is a measure of proximity of $\mathbf{x}_k$ and $\mathbf{x}_\ell$. The caliper width, w, is often taken as a multiple of the standard deviation of the propensity score, $\widehat{e}(\mathbf{x})$, so that by varying

the multiplier, one can vary the relative importance given to $\widehat{e}(\mathbf{x})$ and $\mathbf{x}$. In Table 9.3, the standard deviation of $\widehat{e}(\mathbf{x})$ is 0.172. Tables 9.5 and 9.6 illustrate two distance matrices using a caliper on the propensity score, in which the caliper is half of the standard deviation of the propensity score, or $0.172/2 = 0.086$.

In problems of practical size, a caliper of 20% of the standard deviation of the propensity score is more common, and even that may be too large. A reasonable strategy is to start with a caliper width of 20% of the standard deviation of the propensity score, adjusting the caliper if needed to obtain balance on the propensity score.

In both Tables 9.5 and 9.6, welder #1 has a propensity score that differs from potential controls #1 and #2 by more than 0.086 in absolute value, so the distance is infinite. There is a finite distance between welder #1 and potential controls #3 and #4, because their propensity scores are close, but in both Tables 9.5 and 9.6, the distance is much smaller for potential control #4, because this control is closer to welder #1 in age and smoking behavior. Control #2, who is 63 years old, is at infinite distance from all 21 welders. Experience [38] and simulation [17] suggest that matching on the propensity score $\widehat{e}(\mathbf{x})$ alone often suffices to balance the distribution of covariates, but the individual pairs may be quite different in terms of

Table 9.5 Mahalanobis distances within propensity score calipers. Rows are the 21 welders and columns are for the first 6 of 26 potential controls. An ∞ signifies that the caliper is violated

Welder	Control 1	Control 2	Control 3	Control 4	Control 5	Control 6
1	∞	∞	6.15	0.08	∞	∞
2	∞	∞	∞	0.33	∞	∞
3	∞	∞	∞	∞	∞	∞
4	∞	∞	12.29	∞	∞	∞
5	∞	∞	∞	∞	∞	∞
6	∞	∞	∞	∞	∞	∞
7	∞	∞	∞	∞	∞	∞
8	∞	∞	0.02	5.09	∞	∞
9	∞	∞	∞	∞	∞	∞
10	0.51	∞	∞	∞	10.20	11.17
11	∞	∞	0.18	∞	∞	∞
12	∞	∞	7.06	0.33	∞	∞
13	∞	∞	7.57	∞	∞	∞
14	∞	∞	6.58	0.18	∞	∞
15	∞	∞	∞	∞	∞	∞
16	∞	∞	8.13	∞	∞	∞
17	∞	∞	∞	∞	∞	∞
18	9.41	∞	∞	∞	0.18	0.02
19	∞	∞	∞	∞	∞	∞
20	∞	∞	6.58	0.18	∞	∞
21	∞	∞	∞	∞	∞	∞

Table 9.6 Rank-based Mahalanobis distances within propensity score calipers. Rows are the 21 welders and columns are for the first 6 of 26 potential controls. An ∞ signifies that the caliper is violated

Welder	Control 1	Control 2	Control 3	Control 4	Control 5	Control 6
1	∞	∞	5.98	0.33	∞	∞
2	∞	∞	∞	0.47	∞	∞
3	∞	∞	∞	∞	∞	∞
4	∞	∞	10.43	∞	∞	∞
5	∞	∞	∞	∞	∞	∞
6	∞	∞	∞	∞	∞	∞
7	∞	∞	∞	∞	∞	∞
8	∞	∞	0.04	3.92	∞	∞
9	∞	∞	∞	∞	∞	∞
10	0.25	∞	∞	∞	3.72	4.01
11	∞	∞	0.28	∞	∞	∞
12	∞	∞	7.61	0.98	∞	∞
13	∞	∞	9.02	∞	∞	∞
14	∞	∞	6.83	0.64	∞	∞
15	∞	∞	∞	∞	∞	∞
16	∞	∞	10.61	∞	∞	∞
17	∞	∞	∞	∞	∞	∞
18	3.33	∞	∞	∞	0.05	0.01
19	∞	∞	∞	∞	∞	∞
20	∞	∞	6.83	0.64	∞	∞
21	∞	∞	∞	∞	∞	∞

$\mathbf{x}$, as in the case of welder #1 and potential control #3; whereas minimum distance matching within sufficiently narrow propensity score calipers also tends to balance the distribution of covariates but confers the added benefit of closer individual pairs, as in the case of welder #1 and potential control #4.

In Table 9.5, within the caliper, the distance is the Mahalanobis distance [30, 48], which generalizes to several variables the familiar notion of measuring distance in units of the standard deviation. Speaking very informally, in the Mahalanobis distance, a difference of one standard deviation counts the same for each covariate in $\mathbf{x}$. Even as an informal description, this is not quite correct. The Mahalanobis distance takes account of the correlations among variables. If one covariate in $\mathbf{x}$ were weight in pounds rounded to the nearest pound and another were weight in kilograms rounded to the nearest kilogram, then the Mahalanobis distance would come very close to counting those two covariates as a single covariate because of their high correlation.

If $\widehat{\boldsymbol{\Sigma}}$ is the sample covariance matrix of $\mathbf{x}$, then the estimated Mahalanobis distance [30, 48] between $\mathbf{x}_k$ and $\mathbf{x}_\ell$ is $(\mathbf{x}_k - \mathbf{x}_\ell)^T \, \widehat{\boldsymbol{\Sigma}}^{-1} \, (\mathbf{x}_k - \mathbf{x}_\ell)$. In the welder data in Table 9.3, $\mathbf{x}_\ell = (x_{\ell 1}, x_{\ell 2}, x_{\ell 3})^T$ where $x_{\ell 1}$ is age, $x_{\ell 2}$ is a binary indicator of race,

and $x_{\ell 3}$ is a binary indicator of smoking, and the sample variance covariance matrix is

$$\widehat{\boldsymbol{\Sigma}} = \begin{bmatrix} 54.04 & 0.39 & -0.94 \\ 0.39 & 0.13 & 0.02 \\ -0.94 & 0.02 & 0.25 \end{bmatrix} \tag{9.2}$$

with inverse

$$\widehat{\boldsymbol{\Sigma}}^{-1} = \begin{bmatrix} 0.021 & -0.077 & 0.084 \\ -0.077 & 8.127 & -1.009 \\ 0.084 & -1.009 & 4.407 \end{bmatrix}. \tag{9.3}$$

The Mahalanobis distance was originally developed for use with multivariate Normal data, and for data of that type it works fine. With data that are not Normal, the Mahalanobis distance can exhibit some rather odd behavior. If one covariate contains extreme outliers or has a long-tailed distribution, its standard deviation will be inflated, and the Mahalanobis distance will tend to ignore that covariate in matching. With binary indicators, the variance is largest for events that occur about half the time, and it is smallest for events with probabilities near zero and one. In consequence, the Mahalanobis distance gives greater weight to binary variables with probabilities near zero or one than to binary variables with probabilities closer to one half. In the welder data, two individuals with the same age and race but different smoking behavior have a Mahalanobis distance of 4.407, whereas two individuals with the same age and smoking behavior but different race have a Mahalanobis distance of 8.127, so a mismatch on race is counted as almost twice as bad as a mismatch on smoking. Two people who differed by 20 years in age with the same race and smoking would have a Mahalanobis distance of $0.021 \times 20^2 = 8.4$, so a difference in race counts about as much as a 20-year difference in age. In a different context, if there were binary indicators for the states of the USA, then the Mahalanobis distance would regard matching for Wyoming as vastly more important than matching for California, simply because fewer people live in Wyoming. In many contexts, rare binary covariates are not of overriding importance, and outliers do not make a covariate unimportant, so the Mahalanobis distance may not be appropriate with covariates of this kind.

A simple alternative to the Mahalanobis distance (i) replaces each of the covariates, one at a time, by its ranks, with average ranks for ties, (ii) premultiplies and postmultiplies the covariance matrix of the ranks by a diagonal matrix whose diagonal elements are the ratios of the standard deviation of untied ranks, $1, \ldots, L$, to the standard deviations of the tied ranks of the covariates, and (iii) computes the Mahalanobis distance using the ranks and this adjusted covariance matrix. Call this the "rank-based Mahalanobis distance." Step (i) limits the influence of outliers. After step (ii) is complete, the adjusted covariance matrix has a constant diagonal. Step (ii) prevents heavily tied covariates, such as rare binary variables, from having increased influence due to reduced variance. In the welder data, two

individuals with the same age and race but different smoking behavior have a rank-based Mahalanobis distance of 3.2, whereas two individuals with the same age and smoking behavior but different race have a rank-based Mahalanobis distance of 3.1, so a mismatch on race is counted as about equal to a mismatch on smoking.

Tables 9.5 and 9.6 contrast the Mahalanobis distance and the rank-based Mahalanobis distance for the welder data. In both tables, in row #18 and column #1, there is the distance between a 53-year-old, Caucasian, nonsmoking welder and a 48-year-old, African American, nonsmoking control. In Table 9.5, the distance is 9.41, while in Table 9.6 the distance is 3.33. In fact, the four largest finite distances in Table 9.5 are the four distances between Caucasians and African Americans, whereas race and smoking are about equally important in Table 9.6.

So far, the discussion has used a symmetric caliper on the propensity score. Consider the within pair, treated-minus-control difference in the propensity score. A symmetric caliper views differences of $+0.1$ and -0.1 as equally unfortunate, and makes equal efforts to avoid them. However, if we looked at the 21×26 possible pairs in Table 9.3, positive differences occur more often than negative differences: the propensity score is higher in the treated group, almost by definition. In light of this, perhaps we should view differences in the propensity score asymmetrically, because the bias we are trying to remove is asymmetric. If we paired welder 3 in Table 9.3 to control 21, then the difference in the propensity score would be $0.57 - 0.67 = -0.1$, but perhaps that is better than pairing welder 3 with control 3 with a difference of $.57 - 0.47 = 0.1$, because positive differences are the norm, negative differences the exception. Because we are trying to balance the propensity score, perhaps we should be less tolerant of positive differences, and more tolerant of negative differences. An asymmetric caliper might tolerate, say, any difference between $w/2$ and $-3w/2$ for a fixed $w > 0$, and might penalize any difference outside that interval. Like the symmetric caliper of w to $-w$, the asymmetric caliper of $w/2$ and $-3w/2$ has length $2w$, but it prefers differences that work against the natural direction of the bias. It is possible to show that asymmetric calipers can remove substantially more bias than symmetric calipers [59]; however, the degree of asymmetry must be adjusted to produce the desired effect. In `R`, Ruoqi Yu's `DiPs` package implements asymmetric calipers; see also Sect. 10.5.

In short, a sturdy choice for a distance is the rank-based Mahalanobis distance within calipers on the propensity score, with the caliper width w adjusted to ensure good balance on the propensity score.

9.4 Optimal Pair Matching

An "optimal pair matching" pairs each treated subject with a different control to minimize the total distance within matched pairs [40]. In the welder data in Table 9.3 , this means forming 21 pairs using 21 different controls from the 26 potential controls so that the sum of the 21 distances within pairs is minimized. The

Table 9.7 A small example showing that a greedy algorithm, or a best-first algorithm, does not solve the optimal matching problem. A greedy algorithm would pair treated #1 with control #1, and then would be forced to pair treated #2 with control #2 for a total distance of 0+1000, but it is possible to obtain a match with a total distance of 0.01+0.01

	Control 1	Control 2
Treated 1	0.00	0.01
Treated 2	0.01	1000.00

problem is not trivial because the closest control to one treated subject may also be the closest control to another treated subject. For instance, in Table 9.4, many treated subjects are close to potential control #3, but control #3 will be paired to only one of them. As seen in Table 9.7, a best-first or greedy algorithm will not generally find an optimal pair matching.

Finding an optimal pair matching is known as the "assignment problem" and was solved by Harold Kuhn [28] in 1955. One of the faster algorithms for the assignment problem was proposed by Dimitri Bertsekas [5–7], and Fortran code by Dimitri Bertsekas and Paul Tseng [6] is used by Ben Hansen's [20] `optmatch` package in `R`. Dell'Amico and Toth [15] review the literature on the assignment problem. See also Korte and Vygen [27, Chapter 11].

For the welder data, Table 9.8 displays an optimal match using the propensity score distances in Table 9.4. As the discussion in Sect. 9.3 led us to expect, the marginal means at the bottom of Table 9.8 are fairly well balanced, but the individual pairs are often not matched for race or smoking.

Although it is not obvious from Tables 9.3, 9.5, and 9.6, because of competition for controls, there is no pair matching in which the caliper on the propensity score, $|\widehat{e}(\mathbf{x}_k) - \widehat{e}(\mathbf{x}_\ell)| \leq w$ with $w = 0.086$, is respected for all 21 matched pairs. This is true because of competition among welders for the same controls, despite the fact that each welder has a propensity score within $w = 0.086$ of at least one potential control. For this reason, the infinities in Tables 9.5 and 9.6 are not used; they are replaced by the addition of a "penalty function," which exacts a large but finite penalty for violations of the constraint; see, for instance, [3, pages 372–373] or [26, Chapter 6]. The penalty used here is $1000 \times \max(0, |\widehat{e}(\mathbf{x}_k) - \widehat{e}(\mathbf{x}_\ell)| - w)$, so if $|\widehat{e}(\mathbf{x}_k) - \widehat{e}(\mathbf{x}_\ell)| \leq w$, then the penalty is zero, but if $|\widehat{e}(\mathbf{x}_k) - \widehat{e}(\mathbf{x}_\ell)| > w$, then the penalty is $1000 \times (|\widehat{e}(\mathbf{x}_k) - \widehat{e}(\mathbf{x}_\ell)| - w)$. For instance, the penalty for matching welder #1 to potential control #1 is $1000 \times (|.4587 - 0.1437| - 0.0860) = 229$. This penalty is added to the Mahalanobis distance or rank-based Mahalanobis distance for the corresponding pair. Optimal matching will try to avoid the penalties by respecting the caliper, but when that is not possible, it will prefer to match so the caliper is only slightly violated for a few matched pairs.

Table 9.9 displays the optimal match using the Mahalanobis distances within propensity score calipers from Table 9.5. Again, the calipers are implemented not with ∞'s but rather with a penalty function. Table 9.10 is similar, except the optimal match uses the rank-based Mahalanobis distance within propensity score calipers in Table 9.6, again using the penalty function to implement the calipers. The matches

Table 9.8 Optimal pair match using the squared difference in the propensity score. Covariates are age, race (C = Caucasian, AA = African American), current smoker (Y = yes, N = no), and the estimated propensity score $\widehat{e}(\mathbf{x})$

	Welders				Matched controls			
Pair	Age	Race	Smoker	$\widehat{e}(\mathbf{x})$	Age	Race	Smoker	$\widehat{e}(\mathbf{x})$
1	38	C	N	0.46	45	C	N	0.32
2	44	C	N	0.34	47	C	N	0.28
3	39	C	Y	0.57	39	C	Y	0.57
4	33	AA	Y	0.51	41	C	N	0.40
5	35	C	Y	0.65	34	C	Y	0.67
6	39	C	Y	0.57	31	AA	Y	0.55
7	27	C	N	0.68	35	C	Y	0.65
8	43	C	Y	0.49	41	AA	Y	0.35
9	39	C	Y	0.57	34	C	N	0.54
10	43	AA	N	0.20	50	C	N	0.23
11	41	C	Y	0.53	44	C	Y	0.47
12	36	C	N	0.50	42	C	N	0.38
13	35	C	N	0.52	40	C	N	0.42
14	37	C	N	0.48	44	C	N	0.34
15	39	C	Y	0.57	35	C	N	0.52
16	34	C	N	0.54	38	C	N	0.46
17	35	C	Y	0.65	36	C	Y	0.64
18	53	C	N	0.19	52	C	N	0.20
19	38	C	Y	0.60	36	C	Y	0.64
20	37	C	N	0.48	42	C	N	0.38
21	38	C	Y	0.60	30	C	N	0.63
	Mean	%AA	%Y	Mean	Mean	%AA	%Y	Mean
	38	10	52	0.51	40	10	38	0.46

in Tables 9.9 and 9.10 are similar to each other and better than Table 9.8 in the specific sense that inside individual pairs the matched individuals are closer.

9.5 Optimal Matching with Multiple Controls

In matching with multiple controls, each treated subject is matched to at least one, and possibly more than one, control. To match in a fixed ratio is to match each treated subject to the same number of controls; for instance, pair matching in Sect. 9.4 matches in a ratio of 1-to-1, while matching each treated subject to two controls is matching in a ratio of 1-to-2. To match in a variable ratio is to allow the number of controls to vary from one treated subject to another. In the welder data in Table 9.3 , there are too few potential controls to match in a 1-to-2 fixed ratio, but it is possible to match with multiple controls in a variable ratio, and in particular to use

Table 9.9 Optimal pair match using the Mahalanobis distance within propensity score calipers. The caliper is half the standard deviation of the propensity score, or $0.172/2 = 0.086$. Covariates are age, race (C = Caucasian , AA = African American), current smoker (Y = yes, N = no) and the estimated propensity score $\widehat{e}(\mathbf{x})$

	Welders				Matched controls			
Pair	Age	Race	Smoker	$\widehat{e}(\mathbf{x})$	Age	Race	Smoker	$\widehat{e}(\mathbf{x})$
1	38	C	N	0.46	44	C	N	0.34
2	44	C	N	0.34	47	C	N	0.28
3	39	C	Y	0.57	36	C	Y	0.64
4	33	AA	Y	0.51	41	AA	Y	0.35
5	35	C	Y	0.65	35	C	Y	0.65
6	39	C	Y	0.57	39	C	Y	0.57
7	27	C	N	0.68	30	C	N	0.63
8	43	C	Y	0.49	45	C	N	0.32
9	39	C	Y	0.57	36	C	Y	0.64
10	43	AA	N	0.20	48	AA	N	0.14
11	41	C	Y	0.53	44	C	Y	0.47
12	36	C	N	0.50	41	C	N	0.40
13	35	C	N	0.52	40	C	N	0.42
14	37	C	N	0.48	42	C	N	0.38
15	39	C	Y	0.57	35	C	N	0.52
16	34	C	N	0.54	38	C	N	0.46
17	35	C	Y	0.65	34	C	Y	0.67
18	53	C	N	0.19	52	C	N	0.20
19	38	C	Y	0.60	34	C	N	0.54
20	37	C	N	0.48	42	C	N	0.38
21	38	C	Y	0.60	31	AA	Y	0.55
	Mean	%AA	%Y	Mean	Mean	%AA	%Y	Mean
	38	10	52	0.51	40	14	38	0.45

all of the controls. The decision to match in fixed or variable ratio is a substantial one that affects both the quality of the matching and the analysis and presentation of results. These issues are sketched at the end of this section, following a variable match for the welder data.

Imagine for a moment that there were 42 or more potential controls in the welder data in Table 9.3; then it would be possible to match in a fixed 1-to-2 ratio, producing 21 matched sets, each containing one welder and two controls, where all 42 controls are different. An optimal matching with a fixed ratio of 1-to-2 would minimize the sum of the distances within matched sets between treated subjects and their matched controls, or the sum of 42 distances. Similar considerations apply to matching in a fixed ratio of 1-to-3, 1-to-4, and so on.

Matching with a variable number of controls is slightly more complex. The optimization algorithm decides not only who is matched to whom, but also how many controls to assign to each treated subject. So matching with a variable number

Table 9.10 Optimal pair match using the rank-based Mahalanobis distance within propensity score calipers. The caliper is half the standard deviation of the propensity score, or $0.172/2 = 0.086$. Covariates are age, race (C = Caucasian , AA = African American), current smoker (Y = yes, N = no), and the estimated propensity score $\widehat{e}(\mathbf{x})$

	Welders				Matched controls			
Pair	Age	Race	Smoker	$\widehat{e}(\mathbf{x})$	Age	Race	Smoker	$\widehat{e}(\mathbf{x})$
1	38	C	N	0.46	44	C	N	0.34
2	44	C	N	0.34	47	C	N	0.28
3	39	C	Y	0.57	36	C	Y	0.64
4	33	AA	Y	0.51	41	AA	Y	0.35
5	35	C	Y	0.65	35	C	Y	0.65
6	39	C	Y	0.57	39	C	Y	0.57
7	27	C	N	0.68	30	C	N	0.63
8	43	C	Y	0.49	45	C	N	0.32
9	39	C	Y	0.57	35	C	N	0.52
10	43	AA	N	0.20	48	AA	N	0.14
11	41	C	Y	0.53	44	C	Y	0.47
12	36	C	N	0.50	41	C	N	0.40
13	35	C	N	0.52	40	C	N	0.42
14	37	C	N	0.48	42	C	N	0.38
15	39	C	Y	0.57	36	C	Y	0.64
16	34	C	N	0.54	38	C	N	0.46
17	35	C	Y	0.65	34	C	Y	0.67
18	53	C	N	0.19	52	C	N	0.20
19	38	C	Y	0.60	31	AA	Y	0.55
20	37	C	N	0.48	42	C	N	0.38
21	38	C	Y	0.60	34	C	N	0.54
	Mean	%AA	%Y	Mean	Mean	%AA	%Y	Mean
	38	10	52	0.51	40	14	38	0.45

of controls has more choices to make, but these choices must be constrained in some reasonable way to avoid trivial results. If all the distances were positive, and no constraints were imposed, the minimum distance matching with variable numbers of controls would always be a pair matching. A simple constraint is to insist that a certain number of controls be used. For instance, in the welder data, one might insist that all 26 controls be used. Alternatively, one might insist that 23 controls be used, which would permit, but not require, the algorithm to discard the three potential controls who are older than all of the welders. An attractive feature of fixing the total number of controls is that the total distance is then a sum of a fixed number of distances. In addition to constraining the total number of controls, a different type of constraint permits a treated subject to have, say, at least one but at most three controls. Given some set of constraints on the matching, an optimal matching with variable controls minimizes the total distance between treated subjects and controls inside matched sets.

Table 9.11 Optimal match with multiple controls using the Mahalanobis distance within propensity score calipers. The caliper is half the standard deviation of the propensity score, or $0.172/2 = 0.086$. Covariates are age, race (C = Caucasian, AA = African American), current smoker (Y = yes, N = no) and the estimated propensity score $\widehat{e}(\mathbf{x})$

	Welders				Matched controls			
Matched set	Age	Race	Smoker	$\widehat{e}(\mathbf{x})$	Age	Race	Smoker	$\widehat{e}(\mathbf{x})$
1	38	C	N	0.46	44	C	N	0.34
2	44	C	N	0.34	47	C	N	0.28
3	39	C	Y	0.57	36	C	Y	0.64
4	33	AA	Y	0.51	41	AA	Y	0.35
5	35	C	Y	0.65	35	C	Y	0.65
6	39	C	Y	0.57	36	C	Y	0.64
7	27	C	N	0.68	30	C	N	0.63
8	43	C	Y	0.49	45	C	N	0.32
9	39	C	Y	0.57	35	C	N	0.52
10	43	AA	N	0.20	51	AA	N	0.12
					56	AA	Y	0.13
					48	AA	N	0.14
11	41	C	Y	0.53	44	C	Y	0.47
12	36	C	N	0.50	41	C	N	0.40
13	35	C	N	0.52	40	C	N	0.42
14	37	C	N	0.48	42	C	N	0.38
15	39	C	Y	0.57	39	C	Y	0.57
16	34	C	N	0.54	38	C	N	0.46
17	35	C	Y	0.65	34	C	Y	0.67
18	53	C	N	0.19	63	C	N	0.09
					56	C	N	0.15
					52	C	N	0.20
					50	C	N	0.23
19	38	C	Y	0.60	34	C	N	0.54
20	37	C	N	0.48	42	C	N	0.38
21	38	C	Y	0.60	31	AA	Y	0.55
	Mean	%AA	%Y	Mean	Mean	%AA	%Y	Mean
	38	10	52	0.51	40	14	40	0.45

For the welder data, Table 9.11 is an optimal matching with variable controls, using all 26 controls, based on the Mahalanobis distances with propensity score calipers. The calipers again use the penalty function in Sect. 9.4. The match consists of 19 pairs, one welder matched to three controls, and one welder matched to four controls. Matched set #10 consists of older African Americans. Matched set #18 consists of older Caucasian nonsmokers. Both groups were overrepresented in the control group. As might be expected from the discussion in Sect. 9.3, a variable match using the rank-based Mahalanobis distance (not shown) is generally similar,

but gives about equal emphasis to race and smoking, as opposed to the emphasis on race in Table 9.11.

In a matched pair, a welder is compared with his matched control. For instance, in matched pair #1 in Table 9.11, the welder is 38 years old, the matched control is 44 years old, and the difference in age is $38 - 44 = -6$ years. In matched set #10, the welder is compared with the average of the three matched controls. In this way, the comparison in matched set #10 is just slightly less variable than it would be if one of the three controls were used and the other two were discarded. For instance, in set #10, the welder is 43 years old, the three matched controls have average age $(51 + 56 + 48)/3 = 51.7$, so the difference in age in that set is $43 - 51.7 = -8.67$ years. Similarly, in set #18, the welder is 53 years old, the four matched controls have average age $(63 + 56 + 52 + 50)/4 = 55.25$, so the difference in age in that set is $53 - 55.25 = -2.25$ years. The "average difference in age" is the average of these 21 differences. Similarly, the "average age" in the control group is the average of these 21 average ages of controls. Implicitly, the control in matched set #1 counts as one person, but each of the controls in matched set #10 counts as $\frac{1}{3}$ of a person. In this way, although all controls are used, they are weighted to describe a younger, more Caucasian population with more smokers, that is, a population like the 21 welders. Weighted in this way, the mean age among the matched controls is 40, about two years older than the mean age of 38 for the welders. The same process is applied to a binary variable, such as smoking, to obtain a 40% smoking rate among controls. Although all controls are used in Table 9.11, the balance in the covariate means is similar to pair matching in Sect. 9.4 which discarded five controls. This process of weighting is known as "direct adjustment," and it is one of the basic differences between matching with a fixed ratio and matching with a variable ratio.[3] When matching with a fixed ratio — for instance, with pair matching in Sect. 9.4 — each matched control has the same weight.

The principal advantage of matching in a fixed ratio, such as 1-to-3 matching, is that summary statistics, including those that might be displayed in graphs, may be computed from treated and control groups in the usual way, without using direct adjustment to give unequal weights to observations. For instance, one might display boxplots or Kaplan–Meier survival curves for treated and control groups. This advantage will weigh heavily when potential controls are abundant and the biases that must be removed are not large. A smaller issue is statistical efficiency, which in nominal terms, though often not in practical terms, slightly favors matching in fixed ratio over matching with a variable ratio [32].

[3] Direct adjustment can be applied to pretty much anything, not just means and proportions. In particular, an empirical distribution function is little more than a sequence of proportions, and it is clear how to apply direct adjustment to proportions. From the weighted empirical distribution function, pretty much anything else can be computed; for instance, medians and quartiles. In [21], medians and quartiles from directly adjusted empirical distribution functions are used to construct directly adjusted boxplots. In the weighted empirical distribution function, the 44-year-old control in matched set #1 has mass $1/21$, whereas the 63-year-old control in matched set #18 has mass $1/(4 \times 21)$.

There are several advantages to matching with a variable ratio [32, 40]. First, the matched sets will be more closely matched, in the following precise sense. If one finds the minimum distance match with, say, an *average* of three controls per treated subject and two to four controls for each treated subject, the total distance within matched sets will never be larger, and will typically be quite a bit smaller, than in fixed ratio matching with three controls.[4] This is visible in Table 9.11: the two welders with low propensity scores were matched to the seven controls with low propensity scores, and trying to allocate these seven more evenly among welders would have produced a larger mismatch on these propensity scores. Second, matching in fixed ratio requires the number of controls to be an integer multiple of the number of treated subjects, and this restriction may be inconvenient or undesirable for any of a variety of reasons; for instance, for the welder data in Table 9.3, the only possible matching in fixed ratio is pair matching. Third, just as Sect. 9.1 discussed definite limits to what can be accomplished with pair matching, there are also definite limits to what can be accomplished by matching with multiple controls, but these limits are better when matching with a variable ratio [32].

Finding an optimal match with variable controls is equivalent to solving a particular minimum cost flow problem in a network [40]. In the statistical package `R`, Hansen's [20] `fullmatch` function in his `optmatch` package will determine an optimal match with variable controls from a distance matrix by setting his parameter `min.controls` to equal 1 rather than its default value of 0; the user will often wish to adjust the parameter `max.controls`, which limits the number of controls matched to each treated subject, and the parameter `omit.fraction` which determines the number of controls to use. An alternative approach is to use the assignment algorithm in Sect. 9.4 but with an altered and enlarged distance matrix [33]. For a textbook discussion of network optimization, see [1, 8, 9, 13].

Statistical analysis is not difficult when matching with multiple controls, whether in fixed or variable ratios; see [29, pp. 132–145], [16, pp. 384–387] or [43, pp. 135–139]. For an illustration, in [21, 22], a study of gang violence among teenage boys is analyzed twice, first as a study matched with a fixed ratio of 1-to-2, and second as a study with two strata matched separately, one matched with between two and seven controls and an average of five controls, the other matched with between one and six controls with an average of three controls.

9.6 Optimal Full Matching

In matching with a variable number of controls in Sect. 9.5, a treated subject could be matched to one or more controls. In full matching, the reverse situation

[4] This is true because, when the number of controls matched is the same, optimal matching with variable controls solves a less constrained optimization problem than matching with a fixed ratio, yet the two problems have the same objective function, so the optimum is never worse.

Table 9.12 A distance matrix showing that the best full matching can be much better than the best pair matching or the best matching with variable controls. The best full matching is (1,a,b) and (2,3,c) with a distance of $0.01 + 0.01 + 0.01 + 0.01 = 0.04$. A best pair matching is (1,a), (2,b), (3,c), with a distance of $0.01 + 1000.00 + 0.01 = 1000.02$. Because the number of potential controls equals the number of treated subjects, every matching with variable controls is a pair matching, so variable matching can do no better than pair matching

	Control a	Control b	Control c
Treated 1	0.01	0.01	1000.00
Treated 2	1000.00	1000.00	0.01
Treated 3	1000.00	1000.00	0.01

is permitted as well: one control may be matched to several treated subjects [41]. Table 9.12 shows that full matching can be vastly better than pair matching or matching with a variable number of controls. Because full matching includes as special cases all of the matching procedures considered in Sects. 9.4–9.5, the optimal full matching will produce matched sets that are at least as close as these procedures.

Table 9.13 is an optimal full match for the welder data in Table 9.3, using all 26 controls, and the Mahalanobis distance within propensity score calipers, with the calipers implemented using the penalty function in Sect. 9.4. Matched set #1 is a pair. Matched set #2 has one welder and seven controls. Matched set #3 has one control and seven welders. Within each matched set, the welders and controls look similar.

As in matching with variable controls, summary statistics for the control group in full matching must be directly adjusted. In parallel with Sect. 9.5, the welder in matched set #2 is compared with the average of the seven controls, so these controls implicitly have weight $\frac{1}{7}$. In matched set #3, each of the seven welders is compared with the same control, so this control implicitly has weight 7. As seen at the bottom of Table 9.13, after weighting or direct adjustment, the means in the control group are quite similar to the means in the welder group.

Although the matched sets in Table 9.13 are quite homogeneous, the quite unequal set sizes can lead to some inefficiency. Tables 9.14 and 9.15 are two of many possible variations on the same theme. In Table 9.14, the matched sets can be pairs or triples, and only 21 of the 26 controls are used, in parallel pair matchings in Sect. 9.4. In Table 9.15, all 26 controls are used, with at most two welders or at most three controls in any matched set. In Tables 9.14 and 9.15, welders and controls in the same matched set are reasonably similar, and the (weighted) means in the control group are closer to the welder means than for the several pair matchings in Sect. 9.4.

In one specific sense, an optimal full matching is an optimal design for an observational study [41]. Specifically, define a stratification to be a partitioning of the subjects into groups or strata based on the covariates with the one requirement that each stratum must contain at least one treated subject and at least one control. The quality of a stratification might reasonably be judged by a weighted average of all the within strata distances between treated subjects and controls. For instance,

Table 9.13 Optimal full match using Mahalanobis distance within propensity score calipers. The caliper is half the standard deviation of the propensity score, or $0.172/2 = 0.086$. Covariates are age, race (C = Caucasian, AA = African American), current smoker (Y = yes, N = no), and the estimated propensity score $\widehat{e}(\mathbf{x})$

	Welders				Matched controls			
Matched set	Age	Race	Smoker	$\widehat{e}(\mathbf{x})$	Age	Race	Smoker	$\widehat{e}(\mathbf{x})$
1	38	C	N	0.46	40	C	N	0.42
2	44	C	N	0.34	47	C	N	0.28
					45	C	N	0.32
					44	C	N	0.34
					41	AA	Y	0.35
					42	C	N	0.38
					42	C	N	0.38
					41	C	N	0.40
3	41	C	Y	0.53	39	C	Y	0.57
	39	C	Y	0.57				
	39	C	Y	0.57				
	39	C	Y	0.57				
	39	C	Y	0.57				
	38	C	Y	0.60				
	38	C	Y	0.60				
4	33	AA	Y	0.51	31	AA	Y	0.55
5	35	C	Y	0.65	35	C	Y	0.65
					34	C	Y	0.67
6	27	C	N	0.68	30	C	N	0.63
7	43	C	Y	0.49	44	C	Y	0.47
8	43	AA	N	0.20	51	AA	N	0.12
					56	AA	Y	0.13
					48	AA	N	0.14
9	36	C	N	0.50	35	C	N	0.52
	35	C	N	0.52				
10	37	C	N	0.48	38	C	N	0.46
	37	C	N	0.48				
11	34	C	N	0.54	34	C	N	0.54
12	35	C	Y	0.65	36	C	Y	0.64
					36	C	Y	0.64
13	53	C	N	0.19	63	C	N	0.09
					56	C	N	0.15
					52	C	N	0.20
					50	C	N	0.23
	Mean	%AA	%Y	Mean	Mean	%AA	%Y	Mean
	38	10	52	0.51	39	10	55	0.50

Table 9.14 Optimal full match with sets of size 2 or 3 and 21 controls using the Mahalanobis distance within propensity score calipers. The caliper is half the standard deviation of the propensity score, or $0.172/2 = 0.086$. Covariates are age, race (C = Caucasian, AA = African American), current smoker (Y = yes, N = no), and the estimated propensity score $\widehat{e}(\mathbf{x})$

	Welders				Matched controls			
Matched set	Age	Race	Smoker	$\widehat{e}(\mathbf{x})$	Age	Race	Smoker	$\widehat{e}(\mathbf{x})$
1	38	C	N	0.46	42	C	N	0.38
					42	C	N	0.38
2	44	C	N	0.34	45	C	N	0.32
					44	C	N	0.34
3	39	C	Y	0.57	36	C	Y	0.64
	38	C	Y	0.60				
4	33	AA	Y	0.51	31	AA	Y	0.55
5	35	C	Y	0.65	35	C	Y	0.65
6	39	C	Y	0.57	39	C	Y	0.57
	39	C	Y	0.57				
7	27	C	N	0.68	30	C	N	0.63
8	43	C	Y	0.49	44	C	Y	0.47
	41	C	Y	0.53				
9	43	AA	N	0.20	51	AA	N	0.12
					48	AA	N	0.14
10	36	C	N	0.50	35	C	N	0.52
	35	C	N	0.52				
11	37	C	N	0.48	41	C	N	0.40
					38	C	N	0.46
12	39	C	Y	0.57	36	C	Y	0.64
	38	C	Y	0.60				
13	34	C	N	0.54	34	C	N	0.54
14	35	C	Y	0.65	34	C	Y	0.67
15	53	C	N	0.19	56	C	N	0.15
					52	C	N	0.20
16	37	C	N	0.48	40	C	N	0.42
	Mean	%AA	%Y	Mean	Mean	%AA	%Y	Mean
	38	10	52	0.51	39	10	52	0.50

if a stratum contained two treated subjects and three controls, then the average distance between the treated and control subjects in this stratum is an average of the $6 = 2 \times 3$ distances between two treated and three control subjects. The stratum-specific average distances are combined into a single number, specifically a weighted average distance, perhaps weighting by the number of treated subjects (here 2) or the total number of subjects (here $5 = 2 + 3$) or the number of control subjects (here 3). No matter which of these weightings is used, no matter what distance is used, there is always a full matching that minimizes this weighted average distance [41]. Moreover, with continuous covariates and reasonable distance

Table 9.15 Optimal full match using all 26 controls with sets containing at most two welders and at most three controls, using the Mahalanobis distance within propensity score calipers. The caliper is half the standard deviation of the propensity score, or $0.172/2 = 0.086$. Covariates are age, race (C = Caucasian, AA = African American), current smoker (Y = yes, N = no), and the estimated propensity score $\widehat{e}(\mathbf{x})$

	Welders				Matched controls			
Matched set	Age	Race	Smoker	$\widehat{e}(\mathbf{x})$	Age	Race	Smoker	$\widehat{e}(\mathbf{x})$
1	38	C	N	0.46	44	C	N	0.34
					41	AA	Y	0.35
					42	C	N	0.38
2	44	C	N	0.34	50	C	N	0.23
					47	C	N	0.28
					45	C	N	0.32
3	39	C	Y	0.57	36	C	Y	0.64
	39	C	Y	0.57				
4	33	AA	Y	0.51	31	AA	Y	0.55
5	35	C	Y	0.65	34	C	Y	0.67
6	27	C	N	0.68	30	C	N	0.63
7	43	C	Y	0.49	44	C	Y	0.47
	41	C	Y	0.53				
8	39	C	Y	0.57	39	C	Y	0.57
	39	C	Y	0.57				
9	43	AA	N	0.20	51	AA	N	0.12
					56	AA	Y	0.13
					48	AA	N	0.14
10	36	C	N	0.50	35	C	N	0.52
	35	C	N	0.52				
11	37	C	N	0.48	42	C	N	0.38
					38	C	N	0.46
12	34	C	N	0.54	34	C	N	0.54
13	35	C	Y	0.65	35	C	Y	0.65
14	53	C	N	0.19	63	C	N	0.09
					56	C	N	0.15
					52	C	N	0.20
15	38	C	Y	0.60	36	C	Y	0.64
	38	C	Y	0.60				
16	37	C	N	0.48	41	C	N	0.40
					40	C	N	0.42
	Mean	%AA	%Y	Mean	Mean	%AA	%Y	Mean
	38	10	52	0.51	39	11	56	0.50

functions, with probability 1, only a full matching will minimize the weighted average distance [41]. Stated informally but briefly, a stratification that makes treated subjects and controls as similar as possible is always a full matching.[5]

An optimal full match can be found as a minimum cost flow in a network [19, 41]. In the statistical package `R`, starting with a distance matrix, Hansen's [20] `fullmatch` function in his `optmatch` package will find an optimal full match. Bertsekas and Tseng [6] provide Fortran code for network optimization; this code is called by Hansen's `optmatch` package in `R`. Implementation in `R` is illustrated in Chap. 14.

Full matching is used with an instrument by Hyunseung Kang and colleagues [25]. Peter Austin and Elizabeth Stuart [2] discuss the use of full matching with survival outcomes. See also [57].

9.7 Efficiency

Efficiency is a secondary concern in observational studies; the primary concern is biases that do not diminish as the sample size increases [12]. If there is a bias of fixed size, then as the sample size increases it quickly dominates the mean square error,[6] so one has a highly efficient estimate of the wrong answer. Despite its secondary role, there is an important fact about efficiency with multiple controls that deserves close attention.

Imagine that we will match in fixed ratio, k controls matched to each treated subject in each of I matched sets. Imagine further that the situation satisfies the assumptions associated with the paired t-test. That is: (i) each matched set has a pair parameter that is eliminated by differencing, (ii) there is an additive treatment effect, τ, (iii) beyond the pair effect, there are independent Normally distributed errors with expectation zero, constant variance ω^2. If the matching were perfect, exactly controlling all bias from observed and unobserved covariates, then the treated-minus-control difference in means is unbiased for τ with variance $(1 + 1/k)\,\omega^2/I$. Table 9.16 displays the variance multiplier, $(1 + 1/k)$, for several choices of k. Matching with $k = 2$ substantially reduces the variance of the mean difference, from 2 to 1.5, taking it halfway to $k = \infty$. The gains after that are much smaller, from 1.5 to 1.25 by going from $k = 2$ to $k = 4$, from 1.1 to 1 by going from $k = 10$ to $k = \infty$. See [22] for informal presentation of related results. Detailed efficiency results are found in [21, 32]. If large numbers of close potential controls

[5]Although the proofs of these claims require some attention to detail, the underlying technique may be described briefly. Specifically, if a stratification is not a full match, then some stratum can be subdivided without increasing, and possibly decreasing, its average distance. Subdividing repeatedly terminates in a full match.

[6]There is a sense in which Pitman's asymptotic relative efficiency is not well-defined (or is not meaningful) when there is a bias whose magnitude does not diminish with increasing sample size. See Chap. 16 where this subject is developed in a precise sense.

Table 9.16 Variance ratio $1+1/k$ when matching k controls to each treated subject. Here, $k = \infty$ is only slightly better than $k = 6$, which is slightly better than $k = 4$

Number of controls	1	2	4	6	10	∞
Variance multiplier	2	1.50	1.25	1.17	1.10	1

are available without cost, use of $k = 2$ controls is definitely worthwhile, and $k = 4$ or $k = 6$ controls will often yield some further improvement.

There is abundant evidence from theory [32] and from case studies [56] suggesting that Table 9.16 exaggerates the gain in efficiency from additional controls. The reason is simple: the best $k = 2$ controls will be more closely matched for observed covariates than the best $k = 10$ controls, and the quality of the match affects both bias and variance. The premise of Table 9.16 is that the quality of the match does not change as more controls are used, and that premise is false.

Design sensitivity, discussed in Part III, is the limiting sensitivity to unmeasured bias as the sample size increases. In a precise, quantifiable sense, design sensitivity is far more important than using every available control, and the considerations discussed in Part III may encourage use of some controls and not others; see in particular Chap. 16 and Sect. 18.4. At the same time, with continuous outcomes, increasing the number of controls increases the power of a sensitivity analysis in two ways, by increasing the sample size and by increasing the design sensitivity; see [44] and [45, pp. 222–223].

9.8 Summary

Matching on one variable, the propensity score, tends to produce treated and control groups that, in aggregate, are balanced with respect to observed covariates; however, individual pairs that are close on the propensity score may differ widely on specific covariates. To form closer pairs, a distance is used that penalizes large differences on the propensity score, and then finds individual pairs that are as close as possible. Pairs or matched sets are constructed using an optimization algorithm. Matching with variable controls and full matching combine elements of matching with elements of direct adjustment. Full matching can often produce closer matches than pair matching.

9.9 Further Reading

Elizabeth Stuart [58] reviews matching methods for causal inference. See also [46].

Matching for propensity scores is discussed in [36, 38]. Optimal matching is discussed in [40]. Matching with variable controls is discussed in [32], where it is shown that matching with variable controls can remove more bias than matching

in fixed ratio. The optimality of full matching is proved in [41]. See the paper by Ben Hansen and Stephanie Olsen Klopfer [19] for the best algorithmic results on full matching. For examples of optimal pair matching, see [50, 52–56]; for variable matching, see [21], and for full matching, see the papers by Ben Hansen [18], Elizabeth Stuart and K.M. Green [57], and Ruth Heller, Elisabetta Manduchi, and Dylan Small [23]. For a detailed result on the statistical efficiency of alternative designs, see the appendix of [21].

An attractive place to begin learning about the assignment problem and network optimization is Dimitri Bertsekas' tutorial [7] discussing his auction algorithm. The assignment problem is not trivial because two treated individuals may both want the same control, may both be closest to the same control. The auction algorithm literally holds an auction, selling controls to treated individuals, adjusting the prices of controls as competition emerges for the same control. A treated individual close to several controls may eventually be deterred by the high price of the closest control, opting for a different control at a lower price. The auction algorithm performs well. An additional attraction of the auction algorithm, at least as a place to begin, is that some basic optimization techniques, such as duality and complementary slackness, are placed in an intuitive economic narrative.

9.10 Software

Optimal matching is implemented in `R` by Ben Hansen's [20] `optmatch` package. In addition to Hansen's `optmatch` package, the `match` function in Ruoqi Yu's `DiPs` package also provides a straightforward version of optimal matching for observational studies.

Dimitri Bertsekas and Paul Tseng [6] provide Fortran code for the optimal assignment problem and for network optimization at Bertsekas' website at MIT. Hansen's [20] `pairmatch` function in his `optmatch` package in `R` solves the optimal assignment problem using this Fortran code. The `callrelax` function in Samuel Pimentel's `rcbalance` package in `R` provides a direct interface with the Relax IV Fortran network optimization subroutines of Bertsekas and Tseng [6]; however, note that you must also separately load the `optmatch` package because of its license.

9.11 Data

The data of Costa et al. [14] are in `costa` in the `R` package `DOS2`.

Bibliography

1. Ahuja, R.K., Magnanti, T.L., Orlin, J.B.: Network Flows: Theory, Algorithms, and Applications. Prentice Hall, Upper Saddle River, NJ (1993)
2. Austin, P.C., Stuart, E.A.: Optimal full matching for survival outcomes: a method that merits more widespread use. Stat. Med. **34**, 3949–3967 (2015)
3. Avriel, M.: Nonlinear Programming. Prentice Hall, Upper Saddle River, New Jersey (1976)
4. Bergstralh, E.J., Kosanke, J.L., Jacobsen, S.L.: Software for optimal matching in observational studies. Epidemiology **7**, 331–332 (1996)
5. Bertsekas, D.P.: A new algorithm for the assignment problem. Math. Program. **21**, 152–171 (1981)
6. Bertsekas, D.P., Tseng, P.: The Relax codes for linear minimum cost network flow problems. Ann. Oper. Res. **13**, 125–190 (1988)
7. Bertsekas, D.P.: The auction algorithm for assignment and other network flow problems: A tutorial. Interfaces **20**, 133–149 (1990)
8. Bertsekas, D.P.: Linear Network Optimization. MIT Press, Cambridge, MA (1991)
9. Bertsekas, D.P.: Network Optimization: Continuous and Discrete Models. Athena Scientific, Belmont, MA (1998)
10. Braitman, L.E., Rosenbaum, P.R.: Rare outcomes, common treatments: Analytic strategies using propensity scores. Ann. Intern. Med. **137**, 693–695 (2002)
11. Carpaneto, G., Toth, P.: Algorithm 548: solution of the assignment problem [H]. ACM Trans. Math. Software **6**, 104–111 (1980)
12. Cochran, W.G.: The planning of observational studies of human populations (with discussion). J. R. Stat. Soc. A **128**, 234–265 (1965)
13. Cook, W.J., Cunningham, W.H., Pulleyblank, W.R., Schrijver, A.: Combinatorial Optimization. Wiley, New York (1998)
14. Costa, M., Zhitkovich, A., Toniolo, P.: DNA-protein cross-links in welders: Molecular implications. Cancer Res. **53**, 460–463 (1993)
15. Dell'Amico, M., Toth, P.: Algorithms and codes for dense assignment problems: the state of the art. Discrete Appl. Math. **100**, 17–48 (2000)
16. Fleiss, J.L., Levin, B., Paik, M.C.: Statistical Methods for Rates and Proportions. Wiley, New York (2001)
17. Gu, X.S., Rosenbaum, P.R.: Comparison of multivariate matching methods: Structures, distances, and algorithms. J. Comput. Graph Stat. **2**, 405–420 (1993)
18. Hansen, B.B.: Full matching in an observational study of coaching for the SAT. J. Am. Stat. Assoc. **99**, 609–618 (2004)
19. Hansen, B.B., Klopfer, S.O.: Optimal full matching and related designs via network flows. J. Comput. Graph. Stat. **15**, 609–627 (2006)
20. Hansen, B.B.: Optmatch: Flexible, optimal matching for observational studies. R. News **7**, 18–24 (2007)
21. Haviland, A.M., Nagin, D.S., Rosenbaum, P.R.: Combining propensity score matching and group-based trajectory analysis in an observational study. Psychol. Methods **12**, 247–267 (2007)
22. Haviland, A.M., Nagin, D.S., Rosenbaum, P.R., Tremblay, R.: Combining group-based trajectory modeling and propensity score matching for causal inferences in nonexperimental longitudinal data. Dev. Psychol. **44**, 422–436 (2008)
23. Heller, R., Manduchi, E., Small, D.: Matching methods for observational microarray studies. Bioinformatics **25**, 904–909 (2009)
24. Ho, D., Imai, K., King, G., Stuart, E.A.: Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. Polit. Anal. **15**, 199–236 (2007)
25. Kang, H., Kreuels, B., May, J., Small, D.S.: Full matching approach to instrumental variables estimation with application to the effect of malaria on stunting. Ann. Appl. Stat. **10**, 335–364 (2016)

26. Karmanov, V.G.: Mathematical Programming. Mir, Moscow.
27. Korte, B., Vygen, J.: Combinatorial Optimization, 5th edn. Springer, New York (2012)
28. Kuhn, H.W.: The Hungarian method for the assignment problem. Nav. Res. Logist. Q. **2**, 83–97 (1955)
29. Lehmann, E.L.: Nonparametrics. Holden Day, San Francisco (1975)
30. Mahalanobis, P.C.: On the generalized distance in statistics. Proc. Natl. Inst. Sci. India **12**, 49–55 (1936)
31. Maindonald, J., Braun, J.: Data Analysis and Graphics Using R. Cambridge University Press, New York (2005)
32. Ming, K., Rosenbaum, P.R: Substantial gains in bias reduction from matching with a variable number of controls. Biometrics **56**, 118–124 (2000)
33. Ming, K., Rosenbaum, P.R.: A note on optimal matching with variable controls using the assignment algorithm. J. Comput. Graph. Stat. **10**, 455–463 (2001)
34. R Development Core Team.: R: A Language and Environment for Statistical Computing. R Foundation, Vienna (2007). http://www.R-project.org
35. Papadimitriou, C.H., Steiglitz, K.: Combinatorial Optimization: Algorithms and Complexity. Prentice-Hall, Englewood Cliffs, NJ (1982)
36. Rosenbaum, P.R., Rubin, D.B.: The central role of the propensity score in observational studies for causal effects. Biometrika **70**, 41–55 (1983)
37. Rosenbaum, P.R.: Conditional permutation tests and the propensity score in observational studies. J. Am. Stat. Assoc. **79**, 565–574 (1984)
38. Rosenbaum, P.R., Rubin, D.B.: Constructing a control group by multivariate matched sampling methods that incorporate the propensity score. Am. Stat. **39**, 33–38 (1985)
39. Rosenbaum, P.R.: Model-based direct adjustment. J. Am. Stat. Assoc. **82**, 387–394 (1987)
40. Rosenbaum, P.R.: Optimal matching in observational studies. J. Am. Stat. Assoc. **84**, 1024–1032 (1989)
41. Rosenbaum, P.R.: A characterization of optimal designs for observational studies. J. R. Stat. Soc. B **53**, 597–610 (1991)
42. Rosenbaum, P.R.: Covariance adjustment in randomized experiments and observational studies (with Discussion). Stat. Sci. **17**, 286–327 (2002)
43. Rosenbaum, P.R.: Observational Studies. Springer, New York (2002)
44. Rosenbaum, P.R.: Impact of multiple matched controls on design sensitivity in observational studies. Biometrics **69**, 118–127 (2013)
45. Rosenbaum, P.R.: Observation and Experiment: An Introduction to Causal Inference. Harvard University Press, Cambridge, MA (2017)
46. Rosenbaum, P.R.: Modern algorithms for matching in observational studies. Ann. Rev. Stat. Appl. **7**, 143–176 (2020)
47. Rubin, D.B.: Matching to remove bias in observational studies. Biometrics **29**,159–183 (1973)
48. Rubin, D.B.: Bias reduction using Mahalanobis metric matching. Biometrics **36**, 293–298 (1980)
49. Sekhon, J.S.: Opiates for the matches: Matching methods for causal inference. Annu. Rev. Polit. Sci. **12**, 487–508 (2009)
50. Silber, J.H., Rosenbaum, P.R., Trudeau, M.E., Even-Shoshan, O., Chen, W., Zhang, X., Mosher, R.E.: Multivariate matching and bias reduction in the surgical outcomes study. Med. Care **39**, 1048–1064 (2001)
51. Silber, J.H., Rosenbaum, P.R., Trudeau, M.E., Chen, W., Zhang, X., Lorch, S.L., Rapaport-Kelz, R., Mosher, R.E, Even-Shoshan, O.: Preoperative antibiotics and mortality in the elderly. Ann. Surg. **242**, 107–114 (2005)
52. Silber, J.H., Rosenbaum, P.R., Polsky, D., Ross, R.N., Even-Shoshan, O., Schwartz, S., Armstrong, K.A., Randall, T.C.: Does ovarian cancer treatment and survival differ by the specialty providing chemotherapy? J. Clin. Oncol. **25**, 1169–1175 (2007)
53. Silber, J.H., Lorch, S.L., Rosenbaum, P.R., Medoff-Cooper, B., Bakewell-Sachs, S., Millman, A., Mi, L., Even-Shoshan, O., Escobar, G.E.: Additional maturity at discharge and subsequent health care costs. Health Serv. Res. **44**, 444–463 (2009)

54. Silber, J.H., Rosenbaum, P.R., Clark, A.S., Giantonio, B.J., Ross, R.N., Teng, Y., Wang, M., Niknam, B.A., Ludwig, J.M., Wang, W., Even-Shoshan, O.: Characteristics associated with differences in survival among black and white women with breast cancer. J. Am. Med. Assoc. **310**, 389–397 (2013)
55. Silber, J.H., Rosenbaum, P.R., McHugh, M.D., Ludwig, J.M., Smith, H.L., Niknam, B.A., Even-Shoshan, O., Fleisher, L.A., Kelz, R.R., Aiken, L.H.: Comparison of the value of nursing work environments in hospitals across different levels of patient risk. JAMA Surg. **151**, 527–536 (2016)
56. Smith, H.L.: Matching with multiple controls to estimate treatment effects in observational studies. Sociol. Methodol. **27**, 325–353 (1997)
57. Stuart, E.A., Green, K.M.: Using full matching to estimate causal effects in nonexperimental studies: Examining the relationship between adolescent marijuana use and adult outcomes. Dev. Psychol. **44**, 395–406 (2008)
58. Stuart, E.A.: Matching methods for causal inference. Stat. Sci. **25**, 1–21 (2010)
59. Yu R., Rosenbaum P.R.: Directional penalties for optimal matching in observational studies. Biometrics **75**, 1380–1390 (2019)

Chapter 10
Various Practical Issues in Matching

Abstract Having constructed a matched control group, one must check that it is satisfactory, in the sense of balancing the observed covariates. If some covariates are not balanced, then adjustments are made to bring them into balance. Three adjustments are near-exact matching, exact matching, and the use of small penalties. Exact matching has a special role in extremely large problems, where it can be used to accelerate computation. Matching when some covariates have missing values is discussed.

10.1 Checking Covariate Balance

Although a table such as Table 8.1 in Chap. 8 may suffice to describe covariate balance in a scientific paper [30], somewhat more is typically needed when constructing a matched sample. Checks on covariate balance in matching are informal diagnostics, not unlike residual diagnostics in a regression. They aid in thinking about whether the treated and control groups overlap sufficiently to be matched, and whether the match currently under consideration has achieved reasonable balance or whether some refinements are required. The study of ovarian cancer [30] in Chap. 8 is used to illustrate.

One common measure of covariate imbalance is a slightly unusual version of an absolute standardized difference in means [26]. The numerator of the standardized difference is simply the treated-minus-control difference in covariate means or proportions, and it is computed before and after matching. The first variable, surgeon-type-GO, in Table 8.1 is a binary indicator, which is 1 if a GO performed the surgery and is 0 if someone else performed the surgery. Before matching, the difference in means is $0.76 - 0.33 = 0.43$, or 43%, and after matching it is $0.76 - 0.75 = 0.01$ or 1%. Write $\overline{x}_{tk}$, $\overline{x}_{ck}$, $\overline{x}_{cmk}$ for the means of covariate k in, respectively, the treated group, the control group before matching, and the matched control group, so $\overline{x}_{t1} = 0.76$, $\overline{x}_{c1} = 0.33$, and $\overline{x}_{cmk} = 0.75$ for the first covariate in Table 8.1. It is the denominator that is just slightly unusual in two ways. First, the denominator always describes the standard deviation before matching, even when measuring imbalance after matching. We are asking whether

© Springer Nature Switzerland AG 2020

P. R. Rosenbaum, *Design of Observational Studies*, Springer Series in Statistics,
https://doi.org/10.1007/978-3-030-46405-9_10

the means or proportions are close; we do not want the answer to be hidden by a simultaneous change in the standard deviation. Second, the standard deviation before matching is calculated in a way that gives equal weight to the standard deviation in the treated and control groups before matching. In many problems, the potential control group is much larger than the treated group, but we do not want to give much more weight to the standard deviation in the control group. Write s_{tk} and s_{ck} for the standard deviations of covariate k in the treated group and in the control group before matching. The pooled standard deviation for covariate k is $\sqrt{\left(s_{tk}^2 + s_{ck}^2\right)/2}$. The absolute standardized difference before matching is $sd_{bk} = |\overline{x}_{tk} - \overline{x}_{ck}| / \sqrt{\left(s_{tk}^2 + s_{ck}^2\right)/2}$ and the absolute standardized difference after matching is $sd_{mk} = |\overline{x}_{tk} - \overline{x}_{cmk}| / \sqrt{\left(s_{tk}^2 + s_{ck}^2\right)/2}$; notice that they are identical except that $\overline{x}_{cmk}$ replaces $\overline{x}_{ck}$. For the first covariate in Table 8.1, surgeon-type-GO, the absolute standardized difference is 0.95 before matching and 0.02 after matching, that is, almost a full standard deviation before matching, and about 2% of a standard deviation after matching.

The boxplots in Fig. 10.1 display 67 absolute standardized differences before and after matching. The list of 67 covariates is slightly redundant; for instance, all three categories of surgeon type appear as three binary variables, even though the value of one of these variables is determined by the values of the other two. The imbalances before matching are quite large: there are four covariates with differences of more than half a standard deviation. After matching, the median absolute standardized difference is 0.03 or 3% of a standard deviation, and the maximum is 0.14. In fact, because fine balancing was used to construct this matched sample, 18 of the 67 absolute standardized differences equal zero exactly.

The principal advantage of an absolute standardized difference over an unstandardized difference, say $\overline{x}_{tk} - \overline{x}_{cmk}$, is that variables on different scales, such as age and hypertension, can be plotted in a single graph for quick inspection. The disadvantage is that a covariate such as age means more in terms of years than in terms of standard deviations. In practice, it is helpful to examine an unstandardized table such as Table 8.1 in addition to graphs of standardized differences.

In matching with variable controls, as in Sect. 9.5, or in full matching, as in Sect. 9.6, the mean in the matched control group, $\overline{x}_{cmk}$, is a weighted mean, as described in Sects. 9.5 and 9.6. Then the absolute standardized difference is computed using this weighted mean, $|\overline{x}_{tk} - \overline{x}_{cmk}| / \sqrt{\left(s_{tk}^2 + s_{ck}^2\right)/2}$.

It might seem desirable that all of the absolute standardized differences equal zero, but this would not happen even in a completely randomized experiment. How does the imbalance in the boxplots in Fig. 10.1 compare with the imbalance expected in a completely randomized experiment?

Imagine a completely randomized experiment. This means that 688 unmatched patients are randomly divided into two groups, each with 344 patients. If a randomization test was applied to one covariate to compare the distribution of the covariate in these randomly formed groups, it would produce a P-value less than or equal to 0.05 with probability 0.05; in fact, it would produce a P-value less than

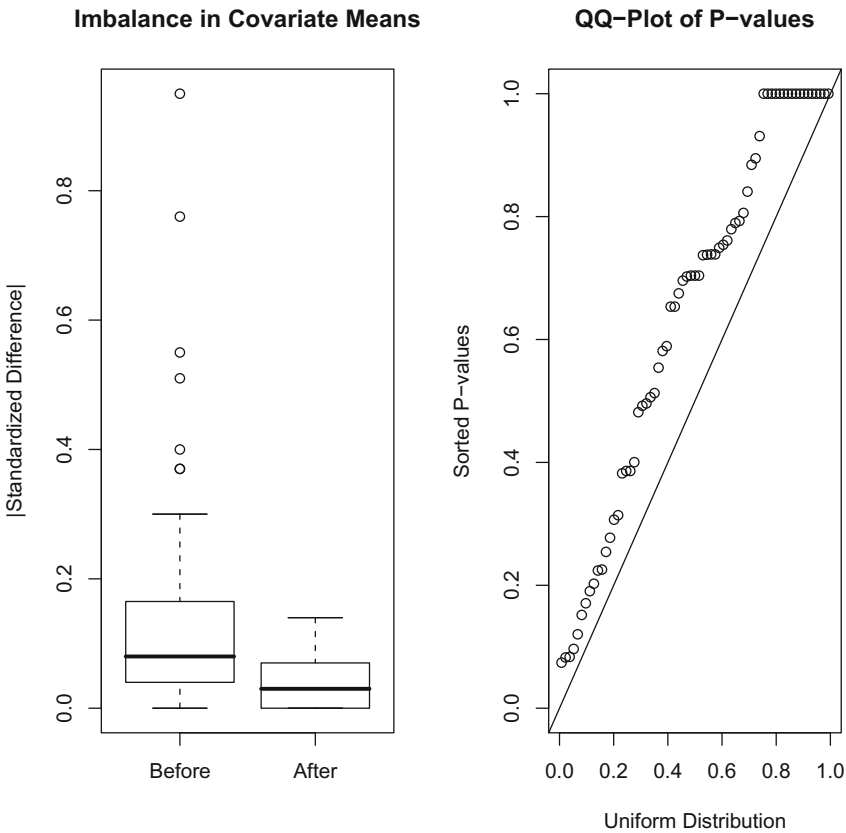

Fig. 10.1 Balance checks for 67 covariates in the study of ovarian cancer. The boxplots show absolute values of standardized differences in means between the GO and MO groups. The quantile–quantile plot compares the 67 two-sample P-values with the uniform distribution, with the line of equality. The boxplot shows that the imbalances in covariates were greatly reduced by matching, while the quantile–quantile plot shows the imbalance on observed covariates after matching is somewhat better than expected for a completely randomized experiment with patients randomly assigned to GO or MO. Balance on observed covariates does not imply balance on covariates that were not observed

or equal to α with probability α for every α between zero and one.[1] Fisher's exact test for a 2×2 table is one such randomization test, and Wilcoxon's rank sum test is another. With 688 patients, it suffices to use the large sample approximations to these tests.

The quantile–quantile plot in Fig. 10.1 compares the 67 two-sample P-values with the uniform distribution. A quantile–quantile plot compares the quantiles in a

[1]This statement is not exactly correct, owing to the discreteness of a randomization distribution, but it is close enough.

sample with the quantiles in a probability distribution to see how they compare [6, pages 143–149]. If the sample looks like the distribution, the points will fall close to the line of equality. In Fig. 10.1 the points fall above the line of equality. This means that the two-sample p-values for the 688 patients tend to be larger than they would be if they came from the uniform distribution. In other words, matching on these 67 covariates has produced more balance on these 67 covariates than we would have expected if we had not matched these patients, but instead had assigned them to treatment or control at random. Of course, randomization has the key benefit that it tends to balance variables that were not measured, whereas matching can only be expected to balance observed covariates.

10.2 Simulated Randomized Experiments as Diagnostics

Randomization tends to balance unobserved covariates, while matching does not. Setting aside that important issue for the moment, we may ask: How does the balance on observed covariates in a matched sample compare with the balance on observed covariates expected in a completely randomized experiment? Section 10.1 addressed this question by applying well-known, established two-sample randomization tests to the observed covariates; see, again, the quantile–quantile plot in Fig. 10.1. However, we need not limit attention to well-known two-sample randomization tests, and there are reasons to also consider other comparisons.

Samuel Pimentel and colleagues [22, §3.2] proposed a simple, general technique measuring balance on observed covariates. Suppose that each of I treated individuals is matched to $K \geq 1$ distinct controls, making I matched sets, each of size $K + 1$, or a total of $I \times (K + 1)$ distinct individuals in the matched sample. The analogous completely randomized experiment picks I of the $I \times (K+1)$ individuals completely at random, making them the treated group, the remaining IK individuals becoming the control group. Taking the observed covariates for the $I \times (K + 1)$ matched individuals, Pimentel and colleagues create this randomized experiment many times, producing one matched observational study and many completely randomized experiments. The marginal distribution of the observed covariates is the same in all of these studies; only the definition of the "treated" group changes. We may then compare the distribution of observed covariates in the matched comparison to the distribution in completely randomized experiments.

In particular, Pimentel and colleagues [22, §3.2] create a single nominal covariate with M levels from the interaction of several nominal covariates. In the extreme case they consider, $M = 176 \times 10^{14}$ or about 2.9 million categories. They form a $2 \times M$ contingency table, treatment-by-covariate, where row one has total count I and row two has total count IK. Were there perfect balance, the count in column m in row two would be K times the count in row one, for $m = 1, \ldots, M$, but even randomization does not produce such perfect balance. In a completely randomized

experiment, this table would have a multivariate hypergeometric distribution, thereby determining the distribution of any measure of imbalance computed from the table. For large M, it is easier to simulate the distribution of a measure of imbalance using the randomized experiments in the previous paragraph.

In their example, Pimentel and colleagues [22, §3.2] considered six nominal covariates, with M ranging from 176 to about 2.9 million. They considered two test statistics: (i) the familiar chi-square statistic for independence in a $2 \times M$ contingency table, and (ii) the total variation distance, essentially the sum over $m = 1, \ldots, M$ of the absolute differences in proportions in the M columns. In their example, in each case, by each measure, the matched sample exhibited better balance than the most balanced of 10,000 randomized experiments built from the same covariates. In effect, this is a P-value of 1 in a test of covariate balance, comparing the matched sample to the distribution of observed covariates in a completely randomized experiment.

In her doctoral thesis, Ruoqi Yu extends this idea. She abandons the $2 \times M$ contingency table, considering instead any functions of the observed covariates as measures of imbalance. For instance, with P covariates, she considers the $P(P-1)/2$ bivariate Kolmogorov–Smirnov statistics [4] applied to pairs of covariates, finding the largest of $P(P-1)/2$ absolute differences in bivariate distributions. That largest absolute difference is compared to its distribution in simulated randomized experiments built from the same covariates. An attractive feature of her approach is that a statistically significant bivariate Kolmogorov–Smirnov statistic suggests an imbalance in a four-category nominal covariate defined by four quadrants for two covariates. That suggested covariate is then added to the match with increased emphasis to bring it into balance, and the balance checks are repeated with the revised match. Her method can identify a covariate imbalance and remove it. The method can be automated and used iteratively. More generally, her method can generate many P-values that check covariate balance, yet also determine the null distribution of the smallest of those many P-values as a correction for testing many hypotheses.

The balance on observed covariates typically produced by complete randomization is a recognizable, intelligible mark on a yardstick that measures covariate imbalance; however, it is nothing more than a mark on a yardstick. By definition, had this degree of imbalance in observed covariates occurred in a completely randomized experiment, we would have judged it unremarkable, and it would not have stirred concern. That fact, by itself, provides no guarantees about causal inference in an observational study.

10.3 Near-Exact Matching

If there are a few covariates of overriding importance, each taking just a few values, then one may wish to match exactly on these covariates whenever this is possible.[2] For instance, in Table 8.1, for reasons discussed in Sect. 8.2, an effort was made to match exactly on both surgeon type and stage, although this was not possible in every instance. The procedure is similar to the use of penalties in Sect. 9.4. If subjects k and ℓ do not have the same values on these few key covariates, then a substantial penalty is added to the distance between them as recorded in the distance matrix. If the penalty is large enough, optimal matching will avoid the penalties whenever possible, and when avoiding all of the penalties is not possible, it will minimize the number of matches that incur the penalty. That is, if an exact match is possible, one will be produced; if not, the match will be as close to exact as possible. In either case, once the match is as exact as possible, the best such match is selected using other considerations represented in the distance matrix. The approach is flexible: it may be used with pair matching (Sect. 9.4), matching with variable controls (Sect. 9.5), or full matching (Sect. 9.6).

Obviously, as the distance matrix increasingly emphasizes these key covariates, it also de-emphasizes the other covariates, perhaps resulting in poor matches on the other covariates. This is the principal disadvantage of all methods that emphasize exact matching: gains in one area are paid for by losses in another, and sometimes the price is high. In contrast, methods that emphasize balancing covariates, such as propensity scores (Sect. 9.2) and fine balance (Chap. 11), may produce gains in one area without losses in another.

To illustrate, in the welder data in Table 9.3, there were 11 smokers among the 21 welders and 9 smokers among the 26 potential controls. In pair matching, an exact match on smoking is not possible: at most 9 of the 11 smoking welders can be matched to smoking controls. In fact, none of the three pair matches in Tables 9.8, 9.9, 9.10 had used all nine smoking controls. Table 10.1 is an optimal pair match using the Mahalanobis distance within propensity score calipers with near-exact matching for smoking. Specifically, the distance matrix for the match in Table 9.5 was used; however, if there was a mismatch on smoking between a welder and a potential control, a penalty was added to the distance between them. This penalty, 5322, was ten times the largest distance in the previous distance matrix used for Table 9.5, where the previous distance matrix already contained penalties to implement the caliper on the propensity score, as described in Sect. 9.4.

In Table 10.1, all nine smoking controls are used, unlike Tables 9.8, 9.9, 9.10. Each of the nine smoking controls is matched to a welder who smokes. Inevitably, two smoking welders are matched to nonsmoking controls; they are #15 and #19. In the means at the bottom of Table 10.1, the balance on smoking is slightly better than in Tables 9.8, 9.9, 9.10, but the balance on race is worse.

[2]Near-exact matching is sometimes called almost-exact matching. These are two phrases with the same meaning.

Table 10.1 Optimal pair match using the Mahalanobis distance within propensity score calipers with near-exact matching for smoking. The caliper is half the standard deviation of the propensity score, or $0.172/2 = 0.086$. Covariates are age, race (C = Caucasian, AA = African American), current smoker (Y = yes, N = no), and the estimated propensity score $\widehat{e}(\mathbf{x})$

	Welders				Matched controls			
Pair	Age	Race	Smoker	$\widehat{e}(\mathbf{x})$	Age	Race	Smoker	$\widehat{e}(\mathbf{x})$
1	38	C	N	0.46	44	C	N	0.34
2	44	C	N	0.34	45	C	N	0.32
3	39	C	Y	0.57	36	C	Y	0.64
4	33	AA	Y	0.51	41	AA	Y	0.35
5	35	C	Y	0.65	35	C	Y	0.65
6	39	C	Y	0.57	39	C	Y	0.57
7	27	C	N	0.68	30	C	N	0.63
8	43	C	Y	0.49	56	AA	Y	0.13
9	39	C	Y	0.57	36	C	Y	0.64
10	43	AA	N	0.20	48	AA	N	0.14
11	41	C	Y	0.53	44	C	Y	0.47
12	36	C	N	0.50	41	C	N	0.40
13	35	C	N	0.52	40	C	N	0.42
14	37	C	N	0.48	42	C	N	0.38
15	39	C	Y	0.57	35	C	N	0.52
16	34	C	N	0.54	38	C	N	0.46
17	35	C	Y	0.65	34	C	Y	0.67
18	53	C	N	0.19	52	C	N	0.20
19	38	C	Y	0.60	34	C	N	0.54
20	37	C	N	0.48	42	C	N	0.38
21	38	C	Y	0.60	31	AA	Y	0.55
	Mean	%AA	%Y	Mean	Mean	%AA	%Y	Mean
	38	10	52	0.51	40	19	43	0.45

Near-exact matching is sometimes useful when there is interest in estimating the treatment effect separately in several subgroups, say for smokers and nonsmokers in Table 10.1. For the ovarian cancer study in Chap. 8, in [30, Table 5], results were reported separately for the 263 pairs in which both patients had stage III or stage IV cancer. Because the matching was "near-exact" for stage, only 10 of 344 pairs had a stage III or IV cancer matched to another stage or missing stage, and so could not be used straightforwardly in such a subgroup analysis.

Near-exact matching adds large penalties to a distance matrix when treated and control subjects differ on a key covariate. Sometimes a covariate is uncomfortably out of balance but not of overriding importance. In this case, small penalties for mismatches are sometimes used. The Mahalanobis distance between two randomly selected individuals from a single multivariate Normal population has expectation equal to twice the number of covariates, so adding a penalty of 2 to a Mahalanobis distance may be viewed, very informally, as doubling the importance of that one covariate.

10.4 Exact Matching

Near-exact matching, as in Sect. 10.3, will result in an exact match whenever one is available, so for many purposes a separate discussion of exact matching is not needed. Nonetheless, there are computational reasons to distinguish exact and near-exact matching.

A simple contingency table will indicate whether an exact match is possible. For the welder data, Table 10.2 shows that exact matching is feasible for race (because $5 \geq 2$ and $21 \geq 19$) but not feasible for smoking (because $9 < 11$), although "near-exact matching" is always feasible, as illustrated in Sect. 10.3.

When feasible, an exact match can be found by subdividing the matching problem into several smaller problems and piecing the answers together. For instance, knowing from Table 10.2 that an exact match for race is feasible, the match can be divided into two separate problems, namely matching the two African American welders and matching the 19 Caucasian welders.

In small problems, there is no advantage to such a subdivision, but in extremely large matching problems, there can be a useful reduction in computational effort, both in required memory storage and in speed of computation. For instance, suppose that there is a nominal covariate with five categories, each containing m treated subjects and n potential controls, so that there are $5m + 5n$ subjects in total. A distance matrix for all $5m+5n$ subjects is $5m \times 5n$ with $25mn$ entries, whereas each of the five subdivided matching problems has a distance matrix that is $m \times n$ with mn entries, so it is $1/25$th as large. The worst-case time bound for the assignment problem grows as the cube of the number of subjects [8, page 147], here $5m + 5n$, so the reduction to a problem of size $m + n$ is expected to substantially accelerate computation.

If a critically important covariate is continuous, it may enter the matching in two versions, one continuous, the other discrete. For instance, a nominal variable can be formed by cutting at quantiles of the continuous variable, and this nominal variable may be used to subdivide the problem for exact matching. The only reasons to do this is to save computer memory or accelerate computation. Within each subdivision, the continuous variable may be used in a distance, such as the Mahalanobis distance, or with a caliper, as described in Sect. 9.4.

Table 10.2 Is exact matching feasible in the welder data? The 2x2 contingency tables show that exact matching is feasible for race, but not for smoking. Covariates are race (C = Caucasian, AA = African American), and current smoker (Y = yes, N = no)

	Smoker	Nonsmoker
Welder	11	10
Potential control	9	17
	AA	C
Welder	2	19
Potential control	5	21

A large sample size should be a luxury, not a hindrance. To enjoy that luxury when matching, consider exact matching on one or two important covariates, subdividing the matching problem into several smaller problems.

10.5 Directional Penalties

A directional penalty pushes against the direction of the bias. Directional penalties are small numbers added to the entries in a distance matrix that act asymmetrically against the natural direction of bias [34]. Directional penalties are often used to correct small imbalances in one or two covariates that are still present after an initial attempt at matching. A few minutes work at the computer adjusting directional penalties can often remove small imperfections in a matched sample that might be difficult to remove by a less focused method.

For instance, by construction, the estimated propensity score $\widehat{e}(\mathbf{x})$ tends to be higher for treated individuals than for potential controls. Matching with a caliper on the propensity score is likely to greatly reduce but not eliminate that tendency. If an initial match exhibits a residual imbalance in the propensity score, rematching with a small directional penalty may remove the imbalance. Suppose that individual k is treated and ℓ is a control, $Z_k = 1$ and $Z_\ell = 0$. A simple directional penalty adds a constant, $\lambda \geq 0$, to the distance between k and ℓ if and only if $\widehat{e}(\mathbf{x}_k) > \widehat{e}(\mathbf{x}_\ell)$. That penalty tends to offset the natural tendency for estimated propensity scores to be higher in the treated group. If two potential controls are equally close to k in terms of the Mahalanobis distance, but one has propensity score $> \widehat{e}(\mathbf{x}_k)$ and the other has a propensity score $< \widehat{e}(\mathbf{x}_k)$, then the penalized distances will slightly favor matching to the first control. The quantity λ may need adjustment and should not be extremely large. Because λ is added to the distance matrix, λ should be large enough to alter but not overwhelm the distances. As discussed at the end of Sect. 10.3, a λ of 2 may be appropriate as a start, subsequently increasing λ if the imbalance remains in the same direction, or reducing λ if the direction of imbalance has been reversed.

The quantity λ is closely related to the use of a Lagrangian to relax a linear constraint in integer programming [34, §2.6]. See [3, §10.3], [10, 19], and [32, §10.1] for discussion of Lagrangians in integer programming.

Directional penalties are also applicable to binary covariates. An alternative version of the technique uses an asymmetric caliper, say requiring $0.05 \geq \widehat{e}(\mathbf{x}_k) - \widehat{e}(\mathbf{x}_\ell) \geq -0.1$ for treated individual k and potential control ℓ. The asymmetric caliper tolerates a larger difference between propensity scores if the difference opposes the natural direction of the bias [34]. As with λ, the degree of asymmetry may be adjusted to produce the desired effect on covariate imbalance.

In `R`, Ruoqi Yu's `DiPs` package implements matching with directional penalties.

10.6 Missing Covariate Values

A covariate has missing values if it is recorded for many people but not for some people. The pattern of missing data refers to which values are missing and which values are present, so you observe the pattern of missing data, though you do not observe the missing values. The constant worry with missing values is that what you cannot see may differ in important ways from what you can see. If a manuscript is missing certain words, it makes all the difference whether the manuscript was damaged by rain or edited by a censor. If edited by a censor, and if you know what motivates the censor, then the fact that a word is missing speaks volumes about what the word must have been, even though you still do not know the word. If a high school student reports his mother's level of education in a survey but not his father's, you might wonder whether he ever met his father. The pattern of missing data may or may not say something about the values that are missing, but knowing the pattern does not tell you the missing values.

The propensity score $e(\mathbf{x})$ is the conditional probability of treatment given the observed covariates, $e(\mathbf{x}) = \Pr(Z = 1 \mid \mathbf{x})$, and matching on $e(\mathbf{x})$ tends to balance $\mathbf{x}$; see Sect. 3.3 or [24, Theorems 1 and 2]. If a coordinate of $\mathbf{x}$ is sometimes missing, think of that coordinate as being a number if it is not missing or * if it is missing. When some covariates have missing values, the propensity score remains well-defined: it refers to the conditional probability of treatment given $\mathbf{x}$, that is $e(\mathbf{x}) = \Pr(Z = 1 \mid \mathbf{x})$, where now some of the coordinates may equal *. It only takes a moment to review the proof of the balancing property of the propensity score to realize that the proof cares not a bit whether the coordinates of $\mathbf{x}$ take only numeric values or sometimes take the value *; see [25, Appendix B] for an explicit and detailed discussion. This says something useful, but alas something less than one might hope. It says that matching on $e(\mathbf{x})$ will tend to balance the observed covariates and the pattern of missing covariates; however, the missing values themselves may or may not be balanced.

The usual way to estimate $e(\mathbf{x})$ when some covariates are missing is to (i) append indicator variables for the pattern of missing data, (ii) for each covariate, plug in an arbitrary but fixed value for each *, and (iii) fit a logit model with these variables, taking the fitted probabilities from the model as the estimates $\widehat{e}(\mathbf{x})$. It is not difficult to verify that the presence of the missing value indicators means that the arbitrary values in step (ii) do not affect $\widehat{e}(\mathbf{x})$, though they do affect the coefficients of the model.[3] Such a logit model is a specific parametric form for a conditional probability $e(\mathbf{x}) = \Pr(Z = 1 \mid \mathbf{x})$ relating two observed quantities, namely Z and $\mathbf{x}$ with its *'s. As is always true with any such model, it may or may not correctly represent the form of $\Pr(Z = 1 \mid \mathbf{x})$—perhaps there should be an interaction linking the missing value indicator for variable #7 with the indicator for variable #19—however, the issues here are no different from any other attempt to

[3]Technically, the fitted probabilities in logit regression are invariant under affine transformations of the predictors.

model a conditional probability of one observable given another. Again, see [25, Appendix B] for additional detail.

Further discussion is aided by an example; see Sect. 14.4.

10.7 Networks and Sparse Representations of Matching

Optimal matching is simply described in terms of a distance matrix, such as Table 9.6. Matrices are familiar, and they fit nicely on a printed page. Nonetheless, Table 9.6 is not an efficient way to represent the information contained in the table. The entry in row i and column j of Table 9.6 is the cost of pairing treated individual i and potential control j, where an infinite distance is intended to forbid this pairing. Most distances in Table 9.6 are infinite, and there is no decision to be made about these potential pairings. Decisions only need to be made when a distance is finite, so a sparse representation of the problem only records the finite distances. When optimal matching is implemented on a computer, much less computer memory is required if the problem is represented to record only finite distances for possible pairings. More efficient use of computer memory permits optimal matching in larger data sets.

An alternative but equivalent representation of optimal matching expresses the problem in terms of a network, that is, a directed graph with various numbers attached to its nodes and edges [2, 27]. Certain edges in the network represent possible pairings, and the absence of such an edge forbids that pairing. There is no edge for an infinite distance in Table 9.6. Exact matching and calipers often imply that most pairings are forbidden, so the network representation is often much smaller than a distance matrix filled with infinities. Software packages for optimal matching in `R` often employ network or other sparse representations of the possible pairings.

In Sect. 10.4, it was noted that exact matching for one or more nominal covariates permits one large problem to be divided into several smaller problems that can be solved much more quickly. Reducing the number of possible pairings—that is, increasing the number of infinite distances in a distance matrix or reducing the number of edges in a network—can have a similar effect, even when the problem cannot be divided in several unrelated problems. If there are T treated individuals and $C \geq T$ potential controls, then the computer time required to find an optimal match can be bounded by $O(C^3)$ as $C \rightarrow \infty$, that is, by $\omega \times C^3$ for some constant ω. However, if as $C \rightarrow \infty$, the number of finite distances in every row remains bounded—say 200 potential controls for each treated individual—then an optimal match may be found much more quickly in computer time bounded by $O\{C^2 \log(C)\}$; see [19]. Obviously, changing the distance matrix or the candidate pairings in the network changes the optimization problem. The point is that sparser problems can be solved more quickly, with the consequence that larger problems can be solved. In `R`, Samuel Pimentel's [22] `rcbalance` package and Ruoqi Yu's [33]

`bigmatch` package accelerate computations by reducing the number of candidate pairings.

It costs something to compute the Mahalanobis distance. We should not compute a distance if we know in advance that we are going to replace it by an infinite distance, or if we are going to omit the corresponding edge from the network. If a caliper forbids matching i to j, then there is no reason to compute the Mahalanobis distance between i and j. In very large matching problems, this elementary step can avoid a substantial amount of unneeded computation [33].

10.8 Constraints on Individual Distances

Optimal matching minimizes the total, or equivalently the mean, distance within matched pairs. A match with a small mean distance within pairs may include a few pairs with comparatively large distances. It is possible to constrain the match to avoid all large distances, and, in the process, to accelerate computations in large data sets [29, 33].

A first thought is to select a threshold, $\kappa > 0$, and to replace by ∞ every distance in the distance matrix that exceeds κ, or equivalently to remove the corresponding edge from the network in Sect. 10.7. This would constrain the maximum distance in a matched pair to be at most κ, and subject to this constraint it would minimize the total or mean distance within pairs. As discussed in Sect. 10.7, fewer edges can produce an algorithm with faster performance in the worst case.

How should the threshold, κ, be selected? If κ is too small, no pair matching will exist. If κ is too large, there may be little reduction in the largest pair differences and little acceleration in the matching algorithm.

One approach determines the smallest κ such that pair matching is feasible. In other words, the match is constrained to minimize the maximum distance, and subject to this constraint, to minimize the total or mean of the within pair distances [29]. This minimax κ may be found using the threshold algorithm of Garfinkel [11]. We select a fairly small initial κ, remove edges with distances larger than κ, and see whether matching is feasible with so few edges. If the initial κ is feasible, we reduce κ and try again. If the initial κ is infeasible, we increase κ and try again. Using binary search, we quickly find a short interval that must contain the minimax κ, and we set κ to the upper endpoint of this interval. Although it is possible to determine the minimax κ with great precision, no practical purpose is served here by many digits of precision in the determination of κ.

An alternative approach is practical even with large administrative databases containing hundreds of thousands of people. Ruoqi Yu and colleagues [33] find the shortest possible caliper on a one-dimensional score, typically the propensity score, and then remove all edges that violate the caliper. Because the score is one-dimensional, a very fast algorithms due to Glover [12, 20] can be used to determine whether matching is feasible with any specified caliper, so that binary search, again, quickly produces a short interval guaranteed to contain the shortest possible caliper.

The method can be adjusted to also incorporate exact matching in Sect. 10.4 and near-fine balance in Chap. 11. In a data set from Medicaid, the example in [33] built, in a single optimization, 38,841 matched pairs from a control reservoir with 159,527 potential controls.

10.9 Clustered Treatment Assignment

In a group-randomized or cluster-randomized trial, individuals are not randomly assigned to treatment or control; rather, naturally occurring clusters of individuals are assigned to treatment or control [9]. It may be practical to provide or withhold a treatment from all children in a school, but impractical to offer it to some children and withhold it from others. The same thing may happen with all patients at the same medical practice. For instance, Bruce et al. [5] randomly assigned medical practices to receive, or not receive, a psychiatric nurse who would aid in identifying depressed patients and ensuring that they received appropriate care. Some covariates may describe individuals, others may describe the clusters that contain the individuals. A randomization inference may need to take account of differences between clusters and differences among individuals within the same cluster, where only the assignment of treatments to clusters is randomized [31]. See also [17].

In parallel, in observational studies, treatments may be assigned to clusters of individuals, now without the benefits of random assignment. José Zubizarreta and Luke Keele [35] proposed an optimal strategy for matching similar schools and similar students within paired schools. They use two steps of dynamic programming. In the first step, for *every* pair of a treated school and a control school, they build the largest balanced matched sample of students, the so-called cardinality match; see Sect. 11.8. Then, in the second step, they pick the best pairs of schools. One consideration in whether to pair two schools in the second step is whether that pairing of schools permits a good pairing of the students within those schools, a consideration that was optimized in the first step. A related method using network optimization was proposed by Samuel Pimentel and colleagues [23].

There is a sense in which a bias that can pick individuals for treatment can do more harm than a bias of similar size that is forced to pick clusters of individuals for treatment. A college admissions officer can reach into every school and offer admission to the best students in each school; that selection process acts on individuals. If a college had to admit whole high schools, rather than pick and choose students within each high school, then it could not confine admission to the best students, and would be forced to admit a mixed group of students, no matter which high schools it admitted. Ben Hansen and colleagues [15] demonstrate this formally: being forced to select clusters for treatments attenuates unmeasured bias, or formally, it increases the design sensitivity.

10.10 Further Reading

For further discussion of checks on covariate balance, see the brief discussion in Cochran's paper [7], and the longer discussions in the papers by Ben Hansen and Jake Bowers [13], Ben Hansen [14], Kosuke Imai, Gary King, and Elizabeth Stuart [16], Sue Marcus et al. [21], and Samuel Pimentel et al. [22]. Propensity scores with missing data are discussed in [25, Appendix B]. The use of penalty functions in optimization is fairly standard; see [1, 18]. Directional penalties are discussed in [34]. Ruoqi Yu and colleagues [33] discuss matching in large administrative databases, as briefly described in Sect. 10.8.

10.11 Software

Ruoqi Yu's `R` package `DiPs` implements matching with directional penalties in Sect. 10.5 with a practical example using data from the US National Health and Nutrition Examination Survery (NHANES). The package also implements exact matching in Sect. 10.4, near-exact matching in Sect. 10.3, and matching with fine or near-fine balance from Chap. 11.

Ruoqi Yu's package `bigmatch` implements her method in [33] and Sect. 10.8 for matching in large administrative databases.

The `R` package `matchMulti` implements the proposal of Pimentel et al. [23] for multilevel matching.

Bibliography

1. Avriel, M.: Nonlinear Programming. Prentice Hall, Englewood Cliffs, NJ (1976)
2. Bertsekas, D.P.: Linear Network Optimization. MIT Press, Cambridge, MA (1991)
3. Bertsekas, D.P.: Network Optimization. Athena Scientific, Belmont, MA (1998)
4. Bickel, P.J.: A distribution free version of the Smirnov two sample test in the p-variate case. Ann. Math. Stat. **40**, 1–23 (1969)
5. Bruce, M.L., Ten Have, T.R., Reynolds, C.F. III, Katz, I.I., Schulberg, H.C., Mulsant, B.H., Brown, G.K., McAvay, G.J., Pearson, J.L., Alexopoulos, G.S.: Reducing suicidal ideation and depressive symptoms in depressed older primary pare patients: a randomized trial. J. Am. Med. Assoc. **291**, 1081–1091 (2004)
6. Cleveland, W.S.: The Elements of Graphing Data. Hobart Press, Summit, NJ (1994)
7. Cochran, W.G.: The planning of observational studies of human populations (with Discussion). J. R. Stat. Soc. A **128**, 234–265 (1965)
8. Cook, W.J., Cunningham, W.H., Pulleyblank, W.R., Schrijver, A.: Combinatorial Optimization. Wiley, New York (1998)
9. Donner, A., Klar, N.: Pitfalls of and controversies in cluster randomization trials. Am. J. Public Health **94**, 416–422 (2004)
10. Fisher, M.L.: The Lagrangian relaxation method for solving integer programming problems. Manag. Sci. **27**, 1–18 (1981)

11. Garfinkel, R.S.: An improved algorithm for the bottleneck assignment problem. Oper. Res. **9**, 1747–1751 (1971)
12. Glover, F.: Maximum matching in a convex bipartite graph. Nav. Res. Logist. Q **14**, 313–316 (1967)
13. Hansen, B.B., Bowers, J.: Covariate balance in simple, stratified and clustered comparative studies. Stat. Sci. **23**, 219–236 (2008)
14. Hansen, B.B.: The essential role of balance tests in propensity-matched observational studies. Stat. Med. **12**, 2050–2054 (2008)
15. Hansen, B.B., Rosenbaum, P.R., Small, D.S.: Clustered treatment assignments and sensitivity to unmeasured biases in observational studies. J. Am. Stat. Assoc. **109**, 133–144 (2014)
16. Imai, K., King, G., Stuart, E.A.: Misunderstandings between experimentalists and observationalists about causal inference. J. R. Stat. Soc. A **171**, 481–502 (2008)
17. Imai, K., King, G., Nall, C.: The essential role of pair matching in cluster-randomized experiments, with application to the Mexican universal health insurance evaluation. Stat. Sci. **24**, 29–53 (2009)
18. Karmanov, V.G.: Mathematical Programming. Mir, Moscow (1989)
19. Korte, B., Vygen, J.: Combinatorial Optimization, 5th edn. Springer, New York (2012)
20. Lipski, W., Preparata, F.P.: Efficient algorithms for finding maximum matchings in convex bipartite graphs and related problems. Acta Informa **15**, 329–346 (1981)
21. Marcus, S.M., Siddique, J., Ten Have, T.R., Gibbons, R.D., Stuart, E., Normand, S-L.T.: Balancing treatment comparisons in longitudinal studies. Psychiatr. Ann. **38**, 12 (2008)
22. Pimentel, S.D., Kelz, R.R., Silber, J.H., Rosenbaum, P.R.: Large, sparse optimal matching with refined covariate balance in an observational study of the health outcomes produced by new surgeons. J. Am. Stat. Assoc. **110**, 515–527 (2015)
23. Pimentel, S.D., Page, L.C., Lenard, M., Keele, L.: Optimal multilevel matching using network flows: an application to a summer reading intervention. Ann. Appl. Stat. **12**, 1479–1505 (2018)
24. Rosenbaum, P.R., Rubin, D.B.: The central role of the propensity score in observational studies for causal effects. Biometrika **70**, 41–55 (1983)
25. Rosenbaum, P.R., Rubin, D.B.: Reducing bias in observational studies using subclassification on the propensity score. J. Am. Stat. Assoc. **79**, 516–524 (1984)
26. Rosenbaum, P.R., Rubin, D.B.: Constructing a control group by multivariate matched sampling methods that incorporate the propensity score. Am. Stat. **39**, 33–38 (1985)
27. Rosenbaum, P.R.: Optimal matching in observational studies. J. Am. Stat. Assoc. **84**, 1024–1032 (1989)
28. Rosenbaum, P.R., Ross R.N., Silber, J.H.: Minimum distance matched sampling with fine balance in an observational study of treatment for ovarian cancer. J. Am. Stat. Assoc. **102**, 75–83. (2007)
29. Rosenbaum, P.R.: Imposing minimax and quantile constraints on optimal matching in observational studies. J. Comput. Graph Stat. **26**, 66–78 (2017)
30. Silber, J.H., Rosenbaum, P.R., Polsky, D., Ross, R.N., Even-Shoshan, O., Schwartz, S., Armstrong, K.A., Randall, T.C.: Does ovarian cancer treatment and survival differ by the specialty providing chemotherapy? J. Clin. Oncol. **25**, 1169–1175 (2007)
31. Small, D.S., Ten Have, T.R., Rosenbaum, P.R.: Randomization inference in a group randomized trial of treatments for depression: covariate adjustment, noncompliance, and quantile effects. J. Am. Stat. Assoc. **103**, 271–279 (2008)
32. Wolsey, L.A.: Integer Programming. Wiley, New York (1998)
33. Yu, R., Silber, J.H., Rosenbaum, P.R.: Matching methods for observational studies derived from large administrative databases. Stat. Sci. (2020, to appear)
34. Yu, R., Rosenbaum, P.R.: Directional penalties for optimal matching in observational studies. Biometrics **75**, 1380–1390 (2019)
35. Zubizarreta, J.R., Keele, L.: Optimal multilevel matching in clustered observational studies: A case study of the effectiveness of private schools under a large-scale voucher system. J. Am. Stat. Assoc. **109**, 547–560 (2017)

Chapter 11
Fine Balance

Abstract Fine balance means constraining a match to balance a nominal variable, without restricting who is matched to whom, when matching to minimize a distance between treated and control subjects. It may be applied to: (1) a nominal variable with many levels that is difficult to balance using propensity scores, (2) a rare binary variable that is difficult to control using a distance, or (3) the interaction of several nominal variables. The fine balance constraint and the distance can emphasize different covariates. When fine balance is unobtainable, near-fine balance is the smallest possible deviation from fine balance.

11.1 What Is Fine Balance?

Fine balance is a constraint on an optimal matching that forces a nominal variable to be balanced [12, 15]. Fine balance is explicitly aimed at covariate balance, and it makes no effort to match exactly for this nominal variable. The pairing can focus on other covariates with the knowledge that this nominal variable will be balanced.

In the study of ovarian cancer in Chap. 8 and [17], the eight SEER sites and three intervals 1991–1992, 1993–1996, and 1997–1999 for year of diagnosis were finely balanced, as was their $24 = 8 \times 3$ category interaction. In other words, the number of patients diagnosed in 1991–1992 in Connecticut, for instance, was exactly the same for GO patients and the matched control group of MO patients, although these patients were not typically matched to each other. As you may recall from Table 8.1, SEER site and year of diagnosis were substantially out of balance prior to matching, but they are not usually viewed as extremely important by clinicians. For this reason, the distance and various penalties used in this match sought individual pairs with the same clinical stage from the same surgeon type with similar propensity scores, and paid no explicit attention to SEER site, yet the fine balance constraint ensured SEER site was balanced in aggregate.

Propensity scores use probability to balance covariates [14], appealing to the law of large numbers [2, Chapter 10]; in effect, if a coin is flipped sufficiently often, the number of heads tends to balance the number of tails. Although they work well in many contexts, propensity scores may not balance covariates when the number

© Springer Nature Switzerland AG 2020

P. R. Rosenbaum, *Design of Observational Studies*, Springer Series in Statistics,
https://doi.org/10.1007/978-3-030-46405-9_11

of trials (or flips) is effectively small. For instance, if the data record little or no information about income, education, pollution, and so on, but do record postal zip code, then one might think to match for zip code as a proxy for various unmeasured variables. In many contexts, there will be only a handful of people from any one zip code, so matching exactly for zip code will yield poor matches for other covariates; moreover, the law of large numbers is of no help. In such a context, fine balancing can balance zip code, without matching people from the same zip code to one another. Similar considerations apply if there are several binary covariates in which one category is rarely seen; these can be combined into one nominal variable with several rare levels and one common level, and this amalgamated covariate can be finely balanced. More generally, if several important nominal variables are thought to interact with each other, fine balancing may be used to exactly balance the many categories formed from their interaction (i.e., their direct product).

In an application of an extension of fine balance—so-called refined balance—Samuel Pimentel et al. [10] balanced nominal covariates with 2.9 million categories. Refined balance is discussed in Sect. 11.6.

11.2 Constructing a Finely Balanced Control Group

As in Chap. 9, fine balance is best understood with the aid of a small example. For the welder data [1] in Chap. 9, it was clear from Tables 9.3 and 10.2 that it was possible to balance race but not smoking. Section 11.2 finds a match that finely balances race, while Sect. 11.3 finds a match that controls the nature of the joint imbalance in race and smoking.

In Tables 9.3 and 10.2, among the potential controls, there are five African Americans where two are needed for balance, and there are 21 Caucasians, where 19 are needed for balance, so three African Americans and two Caucasians must be removed to obtain balance. This is done by expanding the distance matrix, adding five rows. Table 11.1 adds five rows to Table 9.5, which recorded the Mahalanobis distances within propensity score calipers for the first 6 of 26 potential controls. The five extra rows are E1–E5, making the distance matrix square, 26×26. The first two of these extra rows, E1 and E2, will take away two Caucasians, while the last three of the extra rows, E3, E4, and E5, will take away three African Americans. Because control #1 in column 1 of Table 11.1 is African American, he is at infinite distance from the two Caucasian extras and at zero distance from the three African American extras. Because controls #2–#6 in columns 2–6 of Table 11.1 are Caucasian, they are at zero distance from the two Caucasian extras and at infinite distance from the three African American extras. A similar pattern occurs in the remaining 20 columns of the distance matrix.

With a distance matrix patterned as in Table 11.1, an optimal match is found, and the extras and their matched controls are discarded. To avoid the infinite distances, three African American and two Caucasian controls must be matched to the extras, and the match that remains is balanced.

Table 11.1 Mahalanobis distances within propensity score calipers with extra rows to finely balance race

Welder	Control 1	Control 2	Control 3	Control 4	Control 5	Control 6
1	∞	∞	6.15	0.08	∞	∞
2	∞	∞	∞	0.33	∞	∞
3	∞	∞	∞	∞	∞	∞
4	∞	∞	12.29	∞	∞	∞
5	∞	∞	∞	∞	∞	∞
6	∞	∞	∞	∞	∞	∞
7	∞	∞	∞	∞	∞	∞
8	∞	∞	0.02	5.09	∞	∞
9	∞	∞	∞	∞	∞	∞
10	0.51	∞	∞	∞	10.20	11.17
11	∞	∞	0.18	∞	∞	∞
12	∞	∞	7.06	0.33	∞	∞
13	∞	∞	7.57	∞	∞	∞
14	∞	∞	6.58	0.18	∞	∞
15	∞	∞	∞	∞	∞	∞
16	∞	∞	8.13	∞	∞	∞
17	∞	∞	∞	∞	∞	∞
18	9.41	∞	∞	∞	0.18	0.02
19	∞	∞	∞	∞	∞	∞
20	∞	∞	6.58	0.18	∞	∞
21	∞	∞	∞	∞	∞	∞
E1	∞	0	0	0	0	0
E2	∞	0	0	0	0	0
E3	0	∞	∞	∞	∞	∞
E4	0	∞	∞	∞	∞	∞
E5	0	∞	∞	∞	∞	∞

Rows are the 21 welders plus five extras (E), and columns are for the first 6 of 26 potential controls. An ∞ signifies that the caliper is violated

Table 11.2 displays the minimum distance finely balanced match. Race is perfectly balanced, but the pairs are not exactly matched for race. In one sense, Table 11.2 is a slight improvement over Table 9.9 because the control group means for age and smoking are the same, while the means for race and the propensity score are slightly closer to the welder means. Notice that the finely balanced match selected two African American smokers, even though a nonsmoker was available as a control for pair #10.

The general procedure is essentially the same. If there is a nominal covariate for which balance is required, then (1) cross-tabulate that covariate with treatment, as in Table 10.2; (2) determine the number of controls that must be removed from each category of the covariate to achieve balance; (3) add one row for each control that must be removed, with zero distance to its own category and infinite distance

Table 11.2 Optimal pair match using the Mahalanobis distance within propensity score calipers, with fine balance for race

	Welders				Matched controls			
Pair	Age	Race	Smoker	$\widehat{e}(\mathbf{x})$	Age	Race	Smoker	$\widehat{e}(\mathbf{x})$
1	38	C	N	0.46	44	C	N	0.34
2	44	C	N	0.34	47	C	N	0.28
3	39	C	Y	0.57	36	C	Y	0.64
4	33	AA	Y	0.51	41	AA	Y	0.35
5	35	C	Y	0.65	35	C	Y	0.65
6	39	C	Y	0.57	36	C	Y	0.64
7	27	C	N	0.68	30	C	N	0.63
8	43	C	Y	0.49	45	C	N	0.32
9	39	C	Y	0.57	39	C	Y	0.57
10	43	AA	N	0.20	50	C	N	0.23
11	41	C	Y	0.53	44	C	Y	0.47
12	36	C	N	0.50	41	C	N	0.40
13	35	C	N	0.52	40	C	N	0.42
14	37	C	N	0.48	42	C	N	0.38
15	39	C	Y	0.57	35	C	N	0.52
16	34	C	N	0.54	38	C	N	0.46
17	35	C	Y	0.65	34	C	Y	0.67
18	53	C	N	0.19	52	C	N	0.20
19	38	C	Y	0.60	34	C	N	0.54
20	37	C	N	0.48	42	C	N	0.38
21	38	C	Y	0.60	31	AA	Y	0.55
	Mean	%AA	%Y	Mean	Mean	%AA	%Y	Mean
	38	10	52	0.51	40	10	38	0.46

This match is constrained to balance race. The caliper is half the standard deviation of the propensity score, or $0.172/2 = 0.086$. Covariates are age, race (C = Caucasian , AA = African American), current smoker (Y = yes, N = no), and the estimated propensity score $\widehat{e}(\mathbf{x})$

to other categories; (4) find an optimal match for this new, square distance matrix; (5) discard the extra rows and their matched controls. The nominal covariate that is balanced may be formed by combining several other nominal covariates or intervals that cut a continuous variable.

To match two controls to each treated subject with fine balance, stack two copies of the distance matrix, one on top of the other, and then add extras as before to remove imbalances in the nominal covariate. In this way, the number of columns or controls does not change, but each treated subject is represented twice, once in each of two rows. In an optimal match, one control is paired with the first copy of a treated subject and a different control is paired with the second copy, yielding a 2-to-1 match. (Obviously, this cannot be done with the welder data, because at least 42 potential controls are needed, and only 26 are available.) The procedure for

k-to-1 finely balanced matching is similar, but with k copies of the distance matrix rather than two. If needed, a precise algorithmic description is given in [15, §4.2].

It is easy to prove that a minimum distance match with this enlarged distance matrix is a minimum distance match subject to the fine balance constraint, and that the minimum distance match has infinite total distance if and only if the latter problem is infeasible [15, Proposition 1].

The expanded distance matrix in Table 11.1 provides an elementary description of a fast algorithm for minimum distance matching with a fine balance constraint. Although optimal pairing of each row to a different column in the expanded distance matrix does yield fine balance, it is not a concise representation of the problem. As discussed in Sect. 10.7, a more efficient use of computer memory represents matching in terms of a network rather than a distance matrix. The relevant network is discussed in [12, §3.2]. Most current `R` packages that can match with fine balance use a network representation; see Sect. 11.10.

11.3 Controlling Imbalance When Fine Balance Is Not Feasible

When fine balance is not possible, the fine balance procedure can be used to obtain any specified, possible imbalance. The procedure is essentially the same as in Sect. 11.2, except that one must choose an acceptable imbalance from those that are possible.

Table 11.3 displays the options for race and smoking in the welder data. The obstacle to balance on the joint distribution of these two variables is a deficit of smoking, Caucasian controls: there are six, where ten are needed. That is, from the four categories given by the columns of Table 11.3, we would like to select controls with frequencies (1, 10, 1, 9) to agree with the welder group, but the second entry is limited to six, so for a pair match some other entries must increase to compensate. One attractive choice is (3, 6, 0, 12), because this choice uses all nine smoking controls, where 11 would be needed for balance, and is only slightly imbalanced for race, with three African Americans where two would produce balance on race. This choice presumes that smoking is more important than race, as is likely in genetic toxicology, and it provides no information about a treatment-by-race-by-smoking interaction. In any event, for this purely illustrative example, the important

Table 11.3 Table of treatment groups by (race × smoking) for the welder data, with race as C = caucasian , AA = African American, and current smoker coded as Y = yes or N = no

Smoker	Yes	Yes	No	No	
Race	AA	C	AA	C	Total
Welder	1	10	1	9	21
Control	3	6	2	15	26

Table 11.4 Mahalanobis distances within propensity score calipers with extra rows to finely balance race and smoking

Welder	Control 1	Control 2	Control 3	Control 4	Control 5	Control 6
1	∞	∞	6.15	0.08	∞	∞
2	∞	∞	∞	0.33	∞	∞
3	∞	∞	∞	∞	∞	∞
4	∞	∞	12.29	∞	∞	∞
5	∞	∞	∞	∞	∞	∞
6	∞	∞	∞	∞	∞	∞
7	∞	∞	∞	∞	∞	∞
8	∞	∞	0.02	5.09	∞	∞
9	∞	∞	∞	∞	∞	∞
10	0.51	∞	∞	∞	10.20	11.17
11	∞	∞	0.18	∞	∞	∞
12	∞	∞	7.06	0.33	∞	∞
13	∞	∞	7.57	∞	∞	∞
14	∞	∞	6.58	0.18	∞	∞
15	∞	∞	∞	∞	∞	∞
16	∞	∞	8.13	∞	∞	∞
17	∞	∞	∞	∞	∞	∞
18	9.41	∞	∞	∞	0.18	0.02
19	∞	∞	∞	∞	∞	∞
20	∞	∞	6.58	0.18	∞	∞
21	∞	∞	∞	∞	∞	∞
E1	0	∞	∞	∞	∞	∞
E2	0	∞	∞	∞	∞	∞
E3	∞	0	∞	0	0	0
E4	∞	0	∞	0	0	0
E5	∞	0	∞	0	0	0

Rows are the 21 welders plus five extras (E), and columns are for the first 6 of 26 potential controls. An ∞ signifies that the caliper is violated

point that is fine balance can produce any specified possible imbalance, such as (3, 6, 0, 12).

To produce the imbalance (3, 6, 0, 12), one notices that: no controls are to be removed from the first two columns of Table 11.3, two are to be removed from the third column, and three are to be removed from the fourth column. So five rows are added to Table 9.5, two at zero distance from nonsmoking African Americans, three at zero distance from nonsmoking Caucasians, with the other entries set to ∞; see Table 11.4. The first control in Table 9.3 is an African American nonsmoker, and so in column 1 of Table 11.4, he is at zero distance from both of the first two extra rows. The second control in Table 9.3 is a Caucasian nonsmoker, and so, in column 2 of Table 11.4, he is at zero distance from the last three extra rows. The third control in Table 9.3 is a Caucasian smoker, and cannot be eliminated, and so, in column 3 of

Table 11.5 Optimal pair match using the Mahalanobis distance within propensity score calipers, with fine balance for race and smoking

	Welders				Matched controls			
Pair	Age	Race	Smoker	$\widehat{e}(\mathbf{x})$	Age	Race	Smoker	$\widehat{e}(\mathbf{x})$
1	38	C	N	0.46	44	C	N	0.34
2	44	C	N	0.34	47	C	N	0.28
3	39	C	Y	0.57	36	C	Y	0.64
4	33	AA	Y	0.51	41	AA	Y	0.35
5	35	C	Y	0.65	35	C	Y	0.65
6	39	C	Y	0.57	36	C	Y	0.64
7	27	C	N	0.68	30	C	N	0.63
8	43	C	Y	0.49	45	C	N	0.32
9	39	C	Y	0.57	35	C	N	0.52
10	43	AA	N	0.20	56	AA	Y	0.13
11	41	C	Y	0.53	44	C	Y	0.47
12	36	C	N	0.50	41	C	N	0.40
13	35	C	N	0.52	40	C	N	0.42
14	37	C	N	0.48	42	C	N	0.38
15	39	C	Y	0.57	39	C	Y	0.57
16	34	C	N	0.54	38	C	N	0.46
17	35	C	Y	0.65	34	C	Y	0.67
18	53	C	N	0.19	52	C	N	0.20
19	38	C	Y	0.60	31	AA	Y	0.55
20	37	C	N	0.48	42	C	N	0.38
21	38	C	Y	0.60	34	C	N	0.54
	Mean	%AA	%Y	Mean	Mean	%AA	%Y	Mean
	38	10	52	0.51	40	14	43	0.45

The caliper is half the standard deviation of the propensity score, or $0.172/2 = 0.086$. Covariates are age, race (C = Caucasian, AA = African American), current smoker (Y = yes, N = no), and the estimated propensity score $\widehat{e}(\mathbf{x})$. This match is constrained to use all three AA smokers, all six C smokers, neither AA nonsmoker, and 12 C nonsmokers

Table 11.4, he is at infinite distance from all five rows. As in Sect. 11.2, one finds an optimal pair matching in the extended square 26×26 distance matrix and discards the extras and their matched controls.

Table 11.5 is the result. It achieves the prescribed imbalance, $(3, 6, 0, 12)$, and subject to that constraint it minimizes the distance among all pair matchings. Notable in Table 11.5 is the use of all nine smoking controls without a large imbalance in race. This was produced by matching welder #10, a nonsmoking African American, to a smoker, even though a nonsmoking African American control was available.

11.4 Fine Balance, Exact Matching and Near-Exact Matching

Fine balance can exactly balance a nominal covariate without pairing exactly for that covariate. Nonetheless, exact matching for one covariate may be combined with fine balance for another covariate. In parallel with Sect. 10.4, this is one way to implement fine balance in large matching problems: the problem is divided into a few smaller problems by matching exactly for another covariate.

In fact, this was done in the study of ovarian cancer in Chap. 8 and [17]. Three separate matches were constructed, one for each of the three intervals 1991–1992, 1993–1996, and 1997–1999 of year of diagnosis, with fine balance for SEER sites. We did this not primarily for computational reasons but because we had originally envisioned an additional analysis, not reported in [17], that would have compared the speed with which GOs and MOs adopted new chemotherapies for ovarian cancer. The three time intervals 1991–1992, 1993–1996, and 1997–1999 correspond with innovations in chemotherapy for ovarian cancer, and for that analysis it would have been convenient to have pairs matched for the time interval. That is, in each pair, the same chemotherapies would have been available, so it would make sense to ask who used what. Although the matching procedure worked well, we never determined whether GOs or MOs were quicker in adopting new chemotherapies. Instead, we determined that GOs and MOs are both much quicker in adopting new chemotherapies than Medicare is in creating a code for new chemotherapies.

When matching with fine balance is combined with either exact or near-exact matching, a variety of useful side-effects are produced. Suppose, first, that exact matching for one covariate is implemented by splitting the problem into several unrelated subproblems defined by levels of that covariate. Suppose further that a fine balance constraint for another covariate is imposed in each subproblem. Then there is fine balance for interaction (or direct product) of the two covariates. In contrast, if the initial problem is not divided into subproblems and exact matching is imposed by adding a large penalty to the distance matrix, as in Sect. 10.3, then each covariate will be balanced separately, but the interaction of the two covariates will not typically be finely balanced. So these two implementations are quite different, and the former may be infeasible when the latter is feasible. Penalization permits a combination of fine balance for one covariate and near-exact matching for another, so this may be feasible when exact matching is infeasible. If these two covariates are actually the same covariate, then fine balance combined with near-exact matching will attempt to perfectly balance that covariate while matching as often as possible for it; see the example in [23].

11.5 Near-Fine Balance

In Sect. 11.3, fine balance was infeasible, so a specified but feasible imbalance was required. That approach is quite tedious to implement when the nominal covariate has many levels. For instance, there are hundreds of ICD-10 surgical procedure codes, and it would be exceedingly tedious to specify a feasible imbalance for each of them. Instead, this task may be given to the computer: if fine balance is infeasible, find the closest feasible approximation to fine balance. This is matching with near-fine balance, as proposed by Dan Yang et al. [19].

Suppose that a nominal covariate has L levels, $\ell = 1, \ldots, L$. Suppose that there are I treated individuals and each will be matched to one control, so there will be I controls selected from a control reservoir with more than I potential controls. For any possible match, make a $2 \times L$ contingency table, recording treatment, Z, by covariate level for treated individuals and their matched controls; then, each row contains I individuals in total. Fine balance means the count in the first row equals the count in the second, for columns $\ell = 1, \ldots, L$. Fine balance is infeasible if there is no selection of I controls from the control reservoir that yields fine balance; that is, in some columns, we must tolerate a deficit of controls. In a pair match with I treated individuals, we must select I of the potential controls, so a deficit in one category ℓ must be compensated by a surplus in another ℓ'. For each category, ℓ, find the absolute difference between the count for the treated row, $Z = 1$, and the count for the control row, $Z = 0$, and sum these L absolute differences, calling the result the total imbalance. A match exhibits near-fine balance if the total imbalance is as small as possible. If fine balance is feasible, then a total imbalance of zero is possible, and near-fine balance becomes fine balance. If fine balance is infeasible, then the total imbalance cannot be zero, but near-fine balance makes the total imbalance as small as possible. A minimum distance match with near-fine balance minimizes the sum of the I within pair distances subject to the constraint that the total imbalance is minimized.

For instance, no selection of $I = 21$ controls in Table 11.3 produces fine balance, because there is a deficit of four caucasian smoking controls. We can take all six available Caucasian smoking controls, but to have $I = 21$ controls in total, the deficit of four must be distributed among the other columns, so the smallest possible total imbalance is eight. Near-fine balance insists that the total imbalance be eight, and then selects the closest pairing in terms of distance that produces a total imbalance of eight. As defined here, near-fine balance does not care how the surplus is distributed among the other categories. The match we selected in Sect. 11.3 had a larger total imbalance of ten, not eight, because we brought in other considerations besides the total imbalance. Yang et al. [19, §2.1] consider several alternative measures besides the total imbalance.

Near-fine balance may be implemented in any of several ways. The simplest approach, conceptually, is to find a minimum distance pair match in an augmented distance table. In this approach, rows are added, as in Table 11.4, to impose fine balance constraints, but also columns are added to allow controlled deviations from

these fine balance constraints. An added column is at infinite distance from each treated individual, but at zero distance from some of the added rows, permitting such a row to be matched to the extra column rather than to a potential control; see [19, Table 3]. This augmented distance matrix quickly becomes quite large in practice. An equivalent but more concise representation of the problem uses a network rather than a distance matrix, and this network representation is used in many `R` packages; see Sect. 11.10.

11.6 Refined Balance

Near-fine balance in Sect. 11.5 minimizes the total imbalance, the total deviation from fine balance, but it is indifferent to the pattern of the imbalances. In Sect. 11.3, we wanted to balance both smoking and race, but we cared more about smoking, so we were not indifferent to the pattern of the deviations.

Refined balance allows the user to minimize imbalances while controlling their pattern. Instead of one nominal covariate, there is a sequence of nominal covariates, where the next covariate in the sequence subdivides the categories of the current covariate. In Sect. 11.3, the first covariate might be smoking with two categories, while the second covariate might be (smoking $\times$ race) with four categories. Refined covariate balance minimizes the imbalance in the first, coarsest, covariate, and subject to doing that, it minimizes the imbalance in the second covariate in the sequence, and so on. Minimum distance refined balance minimizes this partially ordered, lexically ordered, hierarchy of covariate imbalances, and subject to doing that, it minimizes the total covariate distance within pairs. Refined covariate balance was proposed by Samuel Pimentel et al. [10],

Because the first covariate in the sequence receives priority, adding a second covariate to the hierarchy does not reduce the quality of the balance for the first covariate. Similarly, adding a third covariate does not reduce the quality of the balance for the first two covariates. The example in [10] had as its first covariate 176 surgical procedures, and as its last or sixth covariate was an amalgam with $176 \times 2^{11} = 2.9$ million categories. There were only $I = 6230$ treated individuals. At each of the six levels, the imbalance on the covariate at that level was smaller than the smallest imbalance on that covariate in 10,000 completely randomized experiments; see [10, Table 1] and Sect. 10.2. The match exhibited much better balance on each of the six refined covariates than would be expected from complete randomization.

Minimum distance refined covariate balance is implemented using a network with several layers of balancing nodes [10, Figure 1]; see Sect. 11.10.

11.7 Strength *K* Balance

Instead of balancing one nominal covariate that amalgamates many other simpler nominal covariates, José Zubizarreta [22] proposed balancing marginal distributions of each individual covariate or the joint distribution of each pair of two covariates. By analogy with fractional factorial experiments and orthogonal arrays, a match is said to exhibit strength K balance for several covariates if the joint distribution of any K of these covariates is perfectly balanced [6]. For instance, strength 2 balance is concerned with the joint distributions of covariates two at a time, rather than the joint distribution of all of them at once. Strength K balance is useful when covariates are of comparable importance, lacking the hierarchical structure in Sect. 11.6.

The algorithms needed to produce strength K balance are different from the algorithms used so far in this chapter. As described so far, a minimum distance matching with fine balance, in its various forms, can be found using either as a minimum distance pairing of rows and columns of a distance matrix (the assignment problem) or by a finding a minimum cost flow in a network. There are algorithms that solve these problems relatively quickly; that is, they run in polynomial time. In contrast, strength K balance is produced using mixed integer programming techniques that lack any promise that they will produce a solution in reasonable time. Nonetheless, if the problem is feasible, Zubizarreta's [22] methods typically find a minimum distance matched sample with strength K balance in time that is competitive with network optimization techniques; otherwise, they identify the problem as infeasible. See Sect. 11.10 for discussion of software.

11.8 Cardinality Matching

11.8.1 What Is Cardinality Matching?

Cardinality matching [18, 24] separates two tasks: (1) selecting controls from the control reservoir, and (2) pairing treated individuals with selected controls. In the first step, cardinality matching finds the largest matched sample that meets certain balance requirements. Having completed the first step, cardinality matching may consider various ways of pairing treated individuals and those controls it has selected for the control group in the first step. Cardinality matching makes fine balance the overriding priority; then, pairing is an afterthought.

Balanced distributions do not ensure proximity of individuals. By focusing exclusively on balance in the first step, cardinality matching may miss some opportunities to pair individuals closely. On the other hand, by focusing exclusively on balance in the first step, it may do a better job of balancing distributions.

Cardinality matching is especially useful when the covariates we wish to balance are different from the covariates for which a close pairing is desired. Typically, balance is needed for many covariates. However, we may also want to pair for a few

covariates highly predictive of the outcome. Alternatively, we may want to study effect modification—a treatment effect whose magnitude varies with some observed covariates—and we may want the pairs to be exactly matched for the effect-modifiers. Generally, larger treatment effects are insensitive to larger unmeasured biases; therefore, if there is effect modification, then conclusions are likely to be insensitive to larger biases in some subpopulations defined by observed covariates [5].

11.8.2 Cardinality Matching and Outcome Heterogeneity

José Zubizarreta et al. [24] use cardinality matching to select a control group that balances many covariates. Then, they pair these selected controls to treated individuals using just a few covariates highly predictive of the outcome. Because the second step does not change the matched control group as a group, it changes neither the mean of the treated-minus-control pair differences in outcomes, Y_i, nor the expectation of this mean. However, because the second step pairs closely for covariates highly predictive of the outcome, the second step reduces the variability or heterogeneity of the pair differences, Y_i. As shown in Chap. 16 and [13], reducing heterogeneity of Y_i without altering its expectation tends to make an observational study insensitive to larger biases (i.e., it increases the design sensitivity). José Zubizarreta et al. [24] illustrate this empirically: they pair one control group in two ways, showing that the pairing with less variable Y_i's is less sensitive to unmeasured bias as measured by Γ. To emphasize, the treated and control groups do not change—only the pairing of individuals changes—but because the pair differences, Y_i, are more stable, more consistently positive, the comparison of Y_i with no effect becomes insensitive to a larger bias as measured by Γ.

A different approach that re-pairs matched pairs to reduce heterogeneity of Y_i was proposed by Michael Baiocchi in his doctoral thesis. His method uses Ben Hansen's [4] prognostic score. Unlike the propensity score, which predicts the treatment, the prognostic score predicts the outcome. Unlike the propensity score, it is often best to estimate a prognostic score from an external data set, that is, from data that are not part of the analysis of outcomes. Baiocchi first builds matched pairs using some conventional approach, without using a prognostic score. He then estimates the prognostic score from the controls that were *excluded* from the matched sample. Finally, he re-pairs the existing matched pairs by matching for this external estimate of the prognostic score and, perhaps, for other covariates from the initial match. As above, re-pairing matched pairs does not alter the treated group and the matched control group—it only alters who is matched to whom—so it does not change the covariate balance achieved by the initial match. The hope is that paired individuals who are close on this external estimate of the prognostic score will have similar outcomes, making their pair difference in outcomes, Y_i, more stable, less heterogeneous, and hence more insensitive to unmeasured bias; see Chap. 16 and [13].

11.8.3 Cardinality Matching and Effect Modification

Jesse Hsu et al. [6] use cardinality matching to study effect modification. They first balance many covariates, then try to identify effect-modifiers, then re-pair the control group to be exactly matched for the identified effect-modifiers. If we discover two candidate effect-modifiers among twenty balanced covariates, it is straightforward to re-pair so pairs are extremely close on these two candidates. Of course, covariate balance is unaltered by re-pairing, so the twenty covariates remain balanced. By carefully limiting the scope of the search for effect-modifiers, they pay no price for using the outcomes twice, once to find effect-modifiers, then to analyze the results. Kwonsang Lee et al. [8] apply the method in a study of surgical mortality. See also [5, 7].

11.9 Further Reading

Matching with fine balance is discussed in detail in [15]. The network algorithm in [12, §3.2] will be of interest primarily to programmers because it can be more efficient in its use of space than the procedure described here and in [15]. Refined covariate balance is discussed by Samuel Pimentel et al. [10]. General techniques for balancing covariates were proposed by José Zubizarreta [22].

11.10 Software

In the `R` package `DOS2`, the example in the help-file for the function `fine` reproduces some of the calculations in this chapter. The function `fine` is a companion to this chapter, intended to fix ideas, but I would use one of the packages described below in a practical application of matching with fine balance.

Several `R` packages will find a minimum distance match subject to a constraint of fine balance, near-fine balance, refined balance or strength K balance. The basic function `match` in Ruoqi Yu's package `DiPs` is easy to learn and does matching with fine or near-fine balance. Samuel Pimentel's [9] package `rcbalance` will match with fine, near-fine or refined balance. The `designmatch` package of Cinar Kilcioglu and José Zubizarreta offers a variety of tools for balancing covariates without pairing for those covariates. Ruoqi Yu's `bigmatch` package implements fine and near-fine balance in very large matching problems [20].

11.11 Data

The data of Costa et al. [1] is in `costa` in the `R` package `DOS2`. To try matching with fine balance in a study of practical size, run the example in the help-file for the `match` function in Ruoqi Yu's `DiPs` package in `R`.

Bibliography

1. Costa, M., Zhitkovich, A., Toniolo, P.: DNA-protein cross-links in welders: molecular implications. Cancer Res. **53**, 460–463 (1993)
2. Feller, W.: An Introduction to Probability Theory and Its Applications, vol. 1. Wiley, New York (1968)
3. Hansen, B.B.: Optmatch: flexible, optimal matching for observational studies. R News **7**, 18–24 (2007)
4. Hansen, B.B.: The prognostic analogue of the propensity score. Biometrika **95**, 481–488 (2008)
5. Hsu, J.Y., Small, D.S., Rosenbaum, P.R.: Effect modification and design sensitivity in observational studies. J. Am. Stat. Assoc. **108**, 135–148 (2013)
6. Hsu, J.Y., Zubizarreta, J.R., Small, D.S., Rosenbaum, P.R.: Strong control of the family-wise error rate in observational studies that discover effect modification by exploratory methods. Biometrika **102**, 767–782 (2015)
7. Karmakar, B., Heller, R., Small, D.S.: False discovery rate control for effect modification in observational studies. Elect. J. Stat. **12**, 3232–3253 (2018)
8. Lee, K., Small, D.S., Hsu, J.Y., Silber, J.H., Rosenbaum, P.R.: Discovering effect modification in an observational study of surgical mortality at hospitals with superior nursing. J. R. Stat. Soc. A **181**, 535–546 (2018)
9. Pimentel, S.D.: Large, sparse optimal matching with R package rcbalance. Obs. Stud. **2**, 4–23 (2016)
10. Pimentel, S.D., Kelz, R.R., Silber, J.H., Rosenbaum, P.R.: Large, sparse optimal matching with refined covariate balance in an observational study of the health outcomes produced by new surgeons. J. Am. Stat. Assoc. **110**, 515–527 (2015)
11. Pimentel, S.D., Yoon, F.B., Keele, L.: Variable-ratio matching with fine balance in a study of the peer health exchange. Stat. Med. **34**, 4070–4082 (2015)
12. Rosenbaum, P.R.: Optimal matching in observational studies. J. Am. Stat. Assoc. **84**, 1024–1032 (1989)
13. Rosenbaum, P.R.: Heterogeneity and causality: unit heterogeneity and design sensitivity in observational studies. Am. Stat. **59**, 147–152 (2005)
14. Rosenbaum, P.R., Rubin, D.B.: The central role of the propensity score in observational studies for causal effects. Biometrika **70**, 41–55 (1983)
15. Rosenbaum, P.R., Ross R.N., Silber, J.H.: Minimum distance matched sampling with fine balance in an observational study of treatment for ovarian cancer. J. Am. Stat. Assoc. **102**, 75–83 (2007)
16. Rubin, D.B.: Multivariate matching methods that are equal percent bias reducing, II: maximums on the bias reduction for fixed sample sizes. Biometrics **32**, 121–132 (1976)
17. Silber, J.H., Rosenbaum, P.R., Polsky, D., Ross, R.N., Even-Shoshan, O., Schwartz, S., Armstrong, K.A., Randall, T.C.: Does ovarian cancer treatment and survival differ by the specialty providing chemotherapy? J. Clin. Oncol. **25**, 1169–1175 (2007)
18. Visconti, G., Zubizarreta, J.R.: Handling limited overlap in observational studies with cardinality matching. Obs. Stud. **4**, 217–249 (2018)

19. Yang, D., Small, D.S., Silber, J.H., Rosenbaum, P.R.: Optimal matching with minimal deviation from fine balance in a study of obesity and surgical outcomes. Biometrics **68**, 628–636 (2012)
20. Yu, R., Silber, J.H., Rosenbaum, P.R.: Matching methods for observational studies derived from large administrative databases. Stat. Sci. (2020, to appear)
21. Zaheer, S., Pimentel, S.D., Simmons, K.D., Kuo, L.E., Datta, J., Williams, N., Fraker, D.L., Kelz, R.R.: Comparing international and United States undergraduate medical education and surgical outcomes using a refined balance matching methodology. Ann. Surg. **265**, 916–922 (2017)
22. Zubizarreta, J.R.: Using mixed integer programming for matching in an observational study of kidney failure after surgery. J. Am. Stat. Assoc. **107**, 1360–1371 (2012)
23. Zubizarreta, J.R., Reinke, C., Kelz, R.R., Silber, J.H., Rosenbaum, P.R.: Matching for several sparse nominal variables in a case-control study of readmission following surgery. Am. Stat. **65**, 229–238 (2011)
24. Zubizarreta, J.R., Parades, R.D., Rosenbaum, P.R.: Matching for balance, pairing for heterogeneity in an observational study of the effectiveness of for-profit and not-for-profit high schools in Chile. Ann. App. Stat. **8**, 204–231 (2014)

Chapter 12
Matching Without Groups

Abstract Optimal matching without groups, or optimal nonbipartite matching, offers many additional options for matched designs in both observational studies and experiments. One starts with a square, symmetric distance matrix with one row and one column for each subject recording the distance between any two subjects. Then the subjects are divided into pairs to minimize the total distance within pairs. The method may be used to match with doses of treatment, or with multiple control groups, or as an aid to risk-set matching. An extended discussion of Card and Krueger's study of the minimum wage is used to illustrate.

12.1 Matching Without Groups: Nonbipartite Matching

12.1.1 What Is Nonbipartite Matching?

In previous chapters, treated subjects were matched to controls. That is, there were two groups, treated and control, which existed before the matching began, and members of these two groups were placed in pairs or sets to minimize the distance between treated and control subjects in the same pair or set. The optimization algorithm solved the so-called assignment problem or bipartite matching problem, where "bipartite" makes reference to "two parts." A different optimization problem begins with a single group and divides it into pairs to minimize the total distance between paired units; this is called by the awkward name "nonbipartite" matching. For textbook discussions of the contrast between these two matching problems, see [8, Chapter 5] or [25, Chapter 11]. Nonbipartite matching is highly flexible, and it substantially enlarges the collection of matched designs for observational studies.

In Chap. 9, the distance matrix had one row for each treated subject and one column for each potential control. In contrast, in nonbipartite matching, the distance matrix is square, with one row and one column for every subject. Table 12.1 is a small, artificial example with six subjects, so it is 6×6. The distance matrix is symmetric because the distance from k to ℓ in row k and column ℓ is the same as the distance from ℓ to k in row ℓ and column k. Every subject is at zero distance from itself, but of course a subject cannot be paired with itself.

© Springer Nature Switzerland AG 2020

P. R. Rosenbaum, *Design of Observational Studies*, Springer Series in Statistics,
https://doi.org/10.1007/978-3-030-46405-9_12

Table 12.1 A 6 × 6 distance matrix for nonbipartite matching for six units

ID	1	2	3	4	5	6
1	0	**106**	119	231	110	101
2	**106**	0	207	126	192	68
3	119	207	0	156	247	**25**
4	231	126	156	0	**34**	67
5	110	192	247	**34**	0	212
6	101	68	**25**	67	212	0

Unlike treatment-control matching, every unit appears as both a row and a column of this distance matrix. The optimal nonbipartite matching (1,2), (3,6), (4,5) is shown in **bold** with a minimum total distance of $106 + 25 + 34 = 165$

In Table 12.1, an optimal nonbipartite match would pair the six units into three pairs to minimize the total distance within pairs. The optimal match is shown in bold type: it is (1,2), (3,6), and (4,5) for a total distance of $106 + 25 + 34 = 165$. Notice that subject 1 is closer to subject 6 than to subject 2, but if 1 were matched to 6, then the match for 3 would be much worse. Optimal nonbipartite matching may be done in the statistical package `R`.[1]

12.1.2 Treatment-Control Matching Using an Algorithm for Nonbipartite Matching

Generally, one should not use a nonbipartite matching algorithm to perform bipartite matching because the former requires a larger distance matrix and is somewhat slower. Nonetheless, it is instructive to do it once. Imagine that the first three subjects in Table 12.1 were treated and the last three were controls. To force

[1] Fortran code by Derigs [11] has been made available inside `R` by Bo Lu et al. [24] through a function `nonbimatch(n,d)`, where `d` is the vector form of an `n×n` symmetric matrix of nonnegative integer distances.

```
> dm
     1   2   3   4   5   6
1    0 106 119 231 110 101
2  106   0 207 126 192  68
3  119 207   0 156 247  25
4  231 126 156   0  34  67
5  110 192 247  34   0 212
6  101  68  25  67 212   0
> nonbimatch(6,as.vector(dm))
[1] 2 1 6 5 4 3
```

This says that unit 1 is paired with unit 2, unit 2 is paired with unit 1, unit 3 is paired with unit 6, unit 4 is paired with unit 5, unit 5 is paired with unit 4, and unit 6 is paired with unit 3.

Table 12.2 Using nonbipartite matching to perform treatment-versus-control matching (or bipartite matching)

ID	1	2	3	4	5	6
1	0	∞	∞	231	**110**	101
2	∞	0	∞	**126**	192	68
3	∞	∞	0	156	247	**25**
4	231	**126**	156	0	∞	∞
5	**110**	192	247	∞	0	∞
6	101	68	**25**	∞	∞	0

Subjects 1, 2, and 3 are treated, and infinite distances prevent them from being matched to each other. Subjects 4, 5, and 6 are controls, and infinite distances prevent them from being matched to each other. The optimal match is (1,5), (2,4), (3,6) with distance $110 + 126 + 25 = 261$. Use of a bipartite matching algorithm would have required only a 3×3 distance matrix and be expected to run somewhat more quickly

treated subjects to be matched to controls, infinite distances are placed among treated subjects and among controls, as in Table 12.2. An optimal nonbipartite match in Table 12.2 avoids the infinite distances and pairs (1,5), (2,4), and (3,6), so treated subjects are paired with controls. In this sense, nonbipartite matching is a generalization of bipartite matching. Of course, the bipartite matching could have been performed with a 3×3 distance matrix rather than the 6×6 distance matrix in Table 12.2.

12.1.3 Matching with Doses

Suppose that we do not have treated subjects and controls but rather individuals who received various doses of treatment. In this case, we might wish to form pairs of individuals who are similar in terms of covariates but quite different in terms of doses. As an illustration, imagine that the six individuals 1, 2, ..., 6 in Table 12.1 received the treatment at doses 1, 2, ..., 6, respectively. In this case, we might wish to match so that pairs differ in dose by at least 2; this would preclude matching 1 with 2, for instance.

In Table 12.3, an infinite distance appears whenever two subjects have doses that differ by less than 2. Unlike Table 12.1, the infinite distances constrain who may be matched to whom. Unlike Table 12.2, these constraints are not the sort that can be addressed with a bipartite matching algorithm; that is, the pattern of ∞'s in Table 12.3 does not divide the six units into two nonoverlapping groups. The optimal match in Table 12.3 pairs (1,5), (2,4), and (3,6): that is, this minimizes the distance subject to the constraint that pairs differ in doses by at least 2.

Table 12.3 A 6 × 6 distance matrix for matching with doses of treatment

ID	1	2	3	4	5	6
1	0	∞	119	231	**110**	101
2	∞	0	∞	**126**	192	68
3	119	∞	0	∞	247	**25**
4	231	**126**	∞	0	∞	67
5	**110**	192	247	∞	0	∞
6	101	68	**25**	67	∞	0

For simplicity in this artificial illustration, the doses equal the ID numbers, and the requirement is that paired individuals have doses that differ by at least 2. Where the dose difference is less than 2, the distance is set to ∞. The optimal match is (1,5), (2,4), (3.6)

A practical example of matching with doses was given by Bo Lu, Elaine Zanutto, Robert Hornik, and me [23] in a study of a national media campaign against drug abuse. The dose was the degree of exposure to the media campaign. Another practical example is discussed in Sect. 12.3.

12.1.4 Matching with Several Groups

Suppose that we do not have a treated and control group but rather several groups. In this case, we might wish to pair similar individuals from different groups. For instance, in Table 12.1, suppose that subjects 1 and 2 are in one group, 3 and 4 are in a second group, and 5 and 6 are in a third group. In Table 12.4, individuals in the same group are at infinite distance from one another. The optimal match is (1,3), (2,6), and (4,5). In the optimal match, each individual is paired with an individual from another group to minimize the distance within matched pairs.

The optimal match in Table 12.4 has an additional property. In a balanced incomplete block design, each pair of treatment groups appears together in the same block with the same frequency; see [6, Chapter 11] or [9, §4.2]. The optimal match in Table 12.4 is a very small balanced incomplete block design in the sense that each of the three groups is paired with each of the other groups exactly once. There is a certain equity to a balanced incomplete block design: the number of pairs offering a direct comparison of two groups is the same for every pair of two groups. Bo Lu and I [22] showed how, with three groups, a balanced incomplete block design may be constructed using optimal nonbipartite matching. See [22] for a practical example, and see Sect. 12.2 for a few specifics.

Table 12.4 A 6×6 distance matrix for matching with three groups

ID	1	2	3	4	5	6
1	0	∞	**119**	231	110	101
2	∞	0	207	126	192	**68**
3	**119**	207	0	∞	247	25
4	231	126	∞	0	**34**	67
5	110	192	247	**34**	0	∞
6	101	**68**	25	67	∞	0

Subjects 1 and 2 are in the same group, as are subjects 3 and 4 and subjects 5 and 6. Because subjects in the same group cannot be matched to one another, there is an infinite distance between them. The optimal match is (1,3), (2,6), (4,5) with a total distance of $119 + 68 + 34 = 221$. This is a very small balanced incomplete block design in the sense that every group is paired with every other group exactly once

Table 12.5 Matching with an odd number of subjects

ID	1	2	3	4	5	Sink
1	0	**106**	119	231	110	0
2	**106**	0	207	126	192	0
3	119	207	0	156	247	**0**
4	231	126	156	0	**34**	0
5	110	192	247	**34**	0	0
Sink	0	0	**0**	0	0	0

With five subjects, a sink is added at zero distance to all subjects. One subject is paired with the sink and discarded. In this case, the optimal match is (1,2), (4,5), (3,6) with a distance of $106+34+0 = 140$ and subject 3 is discarded

12.2 Some Practical Aspects of Matching Without Groups

12.2.1 An Odd Number of Subjects

An odd number of subjects cannot be paired. Suppose that we had only the first five subjects in Table 12.1. Then only two pairs may be constructed, with one individual discarded. The best choice of a subject to discard is the one who will make the remaining pairs as close as possible. Beginning with the 5×5 distance matrix for the first five subjects, a "sink" is added at zero distance from all five subjects, making a 6×6 distance matrix; see Table 12.5. When optimal nonbipartite matching is applied to this 6×6 distance matrix, one subject is paired to the sink and discarded. Because any one subject may be discarded at a cost of 0, the matching discards the subject who most improves the total distance for the remaining four subjects within two pairs.

12.2.2 Discarding Some Subjects

Closer matches may be obtained by discarding some subjects. This is done by introducing one "sink" for each subject to be discarded. The sinks are at zero distance from each subject and at infinite distance from one another. The infinite distances prevent sinks from being matched to each other. Optimal nonbipartite matching discards and pairs subjects to minimize the distance within the pairs that remain.

The process is illustrated in Table 12.6. Two sinks remove two subjects to form two pairs. As seen in this small example, discarding the subjects who are most difficult to match can leave behind closely matched pairs.

When there are several groups, as in Table 12.4, one may wish to discard a specific number of subjects from each group, perhaps two subjects from group one, seven subjects from group two, etc. To do this, one introduces sinks in the specified numbers that are at zero distance from the specified groups and at infinite distance from other groups. To discard two subjects from group one, two sinks are introduced at zero distance from subjects in group one and at infinite distance from subjects in other groups. As always, sinks are at infinite distance from one another. In this way, the required subjects in each group are paired with sinks and discarded, and the pairs that remain are as close as possible; see [22] for a proof and an example.

12.2.3 Balanced Incomplete Block Designs with Three Groups

The procedure indicated in Table 12.4 and discussed in Sect. 12.1 suffices to form matched pairs from several groups. In general, the groups will be paired to minimize the distance within pairs, and the pairing may be far from balanced. For instance, group one might always be paired with group two, while group three is always

Table 12.6 Optimal choice of two pairs from six subjects discarding two subjects

ID	1	2	3	4	5	6	Sink	Sink
1	0	106	119	231	110	101	0	**0**
2	106	0	207	126	192	68	**0**	0
3	119	207	0	156	247	**25**	0	0
4	231	126	156	0	**34**	67	0	0
5	110	192	247	**34**	0	212	0	0
6	101	68	**25**	67	212	0	0	0
Sink	0	**0**	0	0	0	0	0	∞
Sink	**0**	0	0	0	0	0	∞	0

The 8×8 distance matrix introduces two sinks at zero distance from each subject and at infinite distance from each other. Two subjects are matched to sinks and discarded. The optimal match is (1,8), (2,7), (3,6), (4,5) with a distance of $0+0+25+34 = 59$. Subjects 1 and 2 are discarded

paired with group four. Whether an unbalanced comparison is undesirable depends upon the context; it may be sensible given the distribution of covariates in the various groups.

In some contexts, one may prefer to force some degree of balance, or even to have a balanced incomplete block design, in which every group is paired equally often with every other group. With three groups, it is possible to control the number of subjects from one group who are paired to subjects from another group, within the limits of arithmetic. Indeed, with the three groups, by simple algebra, the number of subjects used from each group determines the number of subjects in one group paired with another group; see [22, Expression (2.1)]. For instance, in the example in [22, §3.2], 56 subjects were retained in each of three groups, which produced a balanced incomplete block design of minimum distance with 28 pairs of group one with group two, 28 pairs of group one with group three, and 28 pairs of group two with group three.

12.2.4 Propensity Scores for Several Groups

There are several generalizations of the propensity score for use with doses or with several groups. These generalizations depend upon additional assumptions and yield a variety of properties. One generalization uses an ordinal logit model for ordered doses, and matching on the linear portion of this score tends to balance observed covariates when the model is correct; see [18, p. 331]. Scores of this form have been used in nonbipartite matching [23, §2.3]. Another generalization uses several binary logit models as a device for weighting observations [17]. Multivariate models yielding multidimensional scores are also possible; see [18, p. 331] and [16].

12.3 Matching with Doses and Two Control Groups

12.3.1 Does the Minimum Wage Reduce Employment?

The example uses data from David Card and Alan Krueger's [4, 5] study of the effects on employment of increasing the minimum wage; see also Sect. 4.5. It is often said by economists that minimum wage laws harm the people the laws are intended to benefit. For instance, in 1946, George Stigler [33] wrote:

> Each worker receives the value of his marginal product under competition. If a minimum wage is effective, it must therefore have one of two effects: first, workers whose services are worth less than the minimum wage are discharged . . . ; or, second the productivity of low-efficiency workers is increased.

After arguing that the second effect is not plausible, Stigler continues: "The higher the minimum wage, the greater will be the number of covered workers who are

discharged." To my mind, Stigler is discussing a "reason for an effect," in the sense of Sect. 4.6, rather than presenting direct evidence of an effect. For a modern textbook discussion of the minimum wage, see Pierre Cahuc, Stéphane Carcillo, and André Zylberberg [3, §12.2].

In one of the many empirical studies of the effects of the minimum wage on employment, Card and Krueger [4] looked at changes in employment at fast-food restaurants, such as Burger King and Kentucky Fried Chicken, in the adjacent U.S. states of New Jersey and Pennsylvania after New Jersey increased its state minimum wage from \$4.25 to \$5.05 per hour on 1 April 1992. Did the increase in the minimum wage in New Jersey reduce employment at fast-food restaurants?

The increase in the minimum wage in New Jersey from \$4.25 to \$5.05 is expected to have its largest effect on New Jersey restaurants whose starting wage before the increase was at or near the old minimum, namely \$4.25, with little or no effect on restaurants in Pennsylvania where the minimum wage did not increase, and with reduced effect on restaurants in New Jersey having a higher starting wage before the increase. For instance, a New Jersey restaurant paying \$4.25 an hour before the increase would need to raise its starting wage by \$0.80 whereas one paying \$4.75 an hour would comply with the new law by raising its wage by \$0.30; presumably, the latter restaurant is less affected by the law. The theory that the minimum wage depresses employment then yields two predictions, one about a comparison of New Jersey and Pennsylvania, the other about a comparison within New Jersey.

12.3.2 Optimal Matching to Form Two Independent Comparisons

There are 351 restaurants with both starting wages and employment data both before and after the wage increase. They will be divided into 175 pairs of two restaurants, discarding one restaurant, where $351 = 2 \times 175 + 1$. In 65 pairs, there is one Pennsylvania restaurant and one New Jersey restaurant. In the remaining 110 pairs, there are two New Jersey restaurants, one with a low starting wage before the increase, the other with a high starting wage before the increase.

The matching used a 351×351 distance matrix comparing every restaurant to every other restaurant. The distances were a combination of distances on covariates and penalties to force an appropriate comparison. There were five covariates: three binary indicators to distinguish the four restaurant chains (BK = Burger King, KFC = Kentucky Fried Chicken, RR = Roy Rogers, W = Wendy's), a binary indicator of whether or not the restaurant was company owned, and the number of

hours the restaurant was open each day.[2] The rank based Mahalanobis distance was calculated from these five covariates; see Sect. 9.3.

A 66×66 submatrix of the 351×351 distance matrix refers to distances among Pennsylvania restaurants. We do not wish to match Pennsylvania restaurants to one another, so the entries in this 66×66 submatrix were increased by adding a large number, specifically 20 times the largest of the covariate distances, $\mathtt{m} = 37.44$, or $20 \times 37.44 = 748.8$. Any sufficiently large penalty would suffice.

If we match a Pennsylvania restaurant to a New Jersey restaurant, we would like the two matched restaurants to have similar starting wages before the wage increase. A caliper of \$0.20 was used, and it was implemented with a penalty function; see Sect. 9.4. Specifically, if a New Jersey restaurant k and a Pennsylvania restaurant ℓ had starting wages before the increase of w_k and w_ℓ, respectively, then the distance between them was increased by adding the penalty $100 \times \mathtt{m} \times \max(0, |w_k - w_\ell| - 0.2)$. If $|w_k - w_\ell| \leq \$0.20$, the penalty is zero. If $|w_k - w_\ell| = \$0.21$, the penalty is $|w_k - w_\ell| = 100 \times 37.44 \times 0.01 = 37.4$, whereas if $|w_k - w_\ell| = \$0.25$, the penalty is $100 \times 37.44 \times 0.05 = 187.2$. As discussed in Sect. 9.4, the advantage of a graduated penalty function over a single large penalty is that slight violations of the penalty are preferred to large violations. When the matching was completed, among the 65 matched pairs of a restaurant in New Jersey with a restaurant in Pennsylvania, the median difference, $w_k - w_\ell$, NJ−minus−PA, was \$0.00, the median absolute difference, $|w_k - w_\ell|$ was \$0.00, the mean absolute difference was \$0.07, and the maximum absolute difference was \$0.25, so the \$0.20 caliper was violated a few times, but not by much.

If we match a New Jersey restaurant to another New Jersey restaurant, we would like the two matched restaurants to have had very different starting wages before the increase. This too was implemented using a penalty function, with a penalty for a difference less than \$0.50. For two New Jersey restaurants, k and ℓ, with starting wages before the increase of w_k and w_ℓ, the distance between them was increased by adding the penalty $50 \times \max(0,\ 0.5 - |w_k - w_\ell|)$. If $|w_k - w_\ell| \geq \$0.50$, the penalty is zero, but if $|w_k - w_\ell| = \$0.40$, the penalty is 187.2. When the matching

[2]Card and Krueger's [4, 5] data are at Princeton University's Industrial Relations Section webpage http://www.irs.princeton.edu. Restaurants were interviewed twice, once in February 1992, before New Jersey increased its minimum wage, and once in November 1992, after New Jersey increased its minimum wage. Card and Krueger define full-time-equivalent employment as the number of managers plus the number of full-time employees plus half the number of part-time employees, which is NMGRS+EMPFT+EMPPT/2 in their first interview and NMGRS2+EMPFT2+EMPPT2/2 in the second, and the change in employment is the difference between these two quantities, November−minus−February. The price of a full meal refers to the sum of the price of a soda, fries, and an entree, or PSODA+PFRY+PENTREE in the first interview and PSODA2+PFRY2+PENTREE2 in the second, and the change in price is the difference of these two quantities. The starting wage before the increase in the minimum wage is WAGE_ST from the first interview. Other variables used are the restaurant chain (CHAIN), whether the restaurant was company owned (CO_OWNED), and hours open (HRSOPEN). There are 410 restaurants, but the analysis here uses the 351 restaurants with complete data on employment and starting wage. The variable SHEET is Card and Krueger's identification number for a restaurant.

Table 12.7 Distance matrix for six restaurants

Chain	CO	HRS	State	Wage	Sheet	301	310	477	434	208	253
KFC	Yes	10.5	NJ	5.00	301	–	7.4	2067.6	2065.5	233.5	698.0
W	No	11.5	NJ	4.25	310	7.4	–	7.3	4.7	254.8	6.4
BK	No	16.5	PA	4.25	477	2067.6	7.3	–	749.5	641.2	1583.1
BK	No	16.0	PA	4.25	434	2065.5	4.7	749.5	–	642.5	1579.7
RR	Yes	17.0	NJ	4.62	208	233.5	254.8	641.2	642.5	–	475.7
RR	Yes	13.0	NJ	4.87	253	698.0	6.4	1583.1	1579.7	475.7	–

Covariates are: Chain = restaurant chain (BK = Burger King, KFC = Kentucky Fried Chicken, RR = Roy Rogers, W = Wendy's); CO = company owned (Yes, No); and HRS = number of hours open per day. Other variables are State (NJ = New Jersey, PA = Pennsylvania); Wage = starting wage before the wage increase in dollars per hour; and Sheet = Card and Krueger's identification number. The last six columns contain the distances between these six restaurants

was completed, among the 110 pairs of a high-wage and a low-wage New Jersey restaurant, the high−minus−low differences in starting wages before the increase ranged from \$0.25 to \$1.50, with a mean of \$0.58 and a median of \$0.50.

The left side of Table 12.7 shows the baseline information for six of the 351 restaurants, while the right side of Table 12.7 shows the 6×6 portion of the 351×351 distance matrix for these six restaurants. Restaurants with Sheet or identification numbers 301 and 310 are at a distance of 7.4, with no penalties applied, because they are two New Jersey restaurants with very different starting wages before the increase; however, they are not close on the covariates because one is a KFC, the other a Wendy's, one is company owned and the other is not. A large penalty prevents restaurants 301 and 477 from being matched because one is in New Jersey, the other in Pennsylvania, and their starting wages differ by \$0.80. And so on.

An odd number of restaurants cannot be paired. The 351×351 distance matrix is increased in size to 352×352 by adding a row and column of zeros. The added "sink" is at a distance of zero from all actual restaurants. One restaurant—the one restaurant that is hardest to match—is matched to the sink for a cost of zero, and the remaining 350 restaurants are paired into 175 pairs. If 51 sinks were added instead of one sink, with zero distances between restaurants and sinks and infinite distances between one sink and another, then optimal nonbipartite matching would pair 51 restaurants with sinks, leaving 300 restaurants to be paired with each other to form 150 pairs. Use of additional sinks would discard some restaurants, which would be unfortunate, but it would produce closer matches on the pairs that remain. In the current example, one sink was used and one restaurant was discarded.[3]

The optimal nonbipartite matching algorithm was applied to the 352×352 distance matrix to pair the restaurants to minimize the total distance within pairs. Implicitly, the algorithm decides: (1) which one restaurant to discard, (2) which New Jersey restaurants should be paired with Pennsylvania restaurants and which

[3] As it turned out, the discarded restaurant was a company owned Pennsylvania KFC open 10 h per day and paying a starting wage before the increase of \$5.25. It was Sheet #481.

Table 12.8 Balance on one covariate, hours open per day, in All 175 pairs of a more affected and a less affected restaurant, in 65 NJ-vs.-PA pairs, and in 110 NJ-vs.-NJ pairs

Group	Type	Minimum	Lower quartile	Median	Upper quartile	Maximum
More affected	All	7.0	11.5	15.5	16.0	24.0
Less affected	All	8.0	12.0	16.0	16.5	24.0
More affected	NJ-vs.-PA	7.0	11.5	16.0	16.5	19.0
Less affected	NJ-vs.-PA	8.0	12.0	16.0	16.5	24.0
More affected	NJ-vs.-NJ	10.0	11.5	15.0	16.0	24.0
Less affected	NJ-vs.-NJ	9.5	12.0	15.0	16.0	24.0

In addition to this one covariate, the pairs were also matched for restaurant chain, with 88% matched exactly, and for company owned, with 93% matched exactly. The NJ-vs.-PA pairs had similar starting wages before the increase in the NJ minimum wage. The NJ-vs.-NJ pairs had quite different starting wages before the increase, the less affected restaurant having a higher starting wage

should be paired with other New Jersey restaurants, (3) which individual restaurants should be paired.

For the optimal match, Table 12.8 describes the covariates. In Table 12.8, "more affected" refers to the New Jersey restaurant in a NJ-vs.-PA pair, and it refers to the restaurant with the lower starting wage before the increase in a NJ-vs.-NJ pair. Of course, "less affected" refers to the other restaurant in the same pair. In 88% of the pairs, the restaurants were the same restaurant chain, and in 93% of the pairs, they were the same in terms of company ownership. In addition, the distribution of hours open was similar, as seen in Table 12.8.

For the optimal match, Fig. 12.1 and Table 12.9 show the distribution of starting wages before the wage increase. As was intended, the distribution of starting wages was quite similar in the 65 NJ-vs.-PA pairs, and it was very different in the 110 NJ-vs.-NJ pairs. On average, the more affected restaurant in a NJ-vs.-NJ pair would have to raise the starting wage by \$0.73 to comply with the increase in the minimum wage to \$5.05, whereas the less affected restaurant would need an increase of just \$0.14.

12.3.3 Difference in Change in Employment with Two Control Groups

Did the increase in New Jersey's minimum wage reduce employment in more affected restaurants in the 175 pairs? The outcome is the change in full-time-equivalent employment (FTE) following the wage increase, after−minus−before. In the discussion that follows, a "worker" or an "employee" refers to one FTE worker, even if that means two people each working half time. The typical restaurant had about 20 workers. A change of −1 means one less worker in a restaurant after the wage increase. Table 12.10 and Fig. 12.2 display the distribution of changes. A few

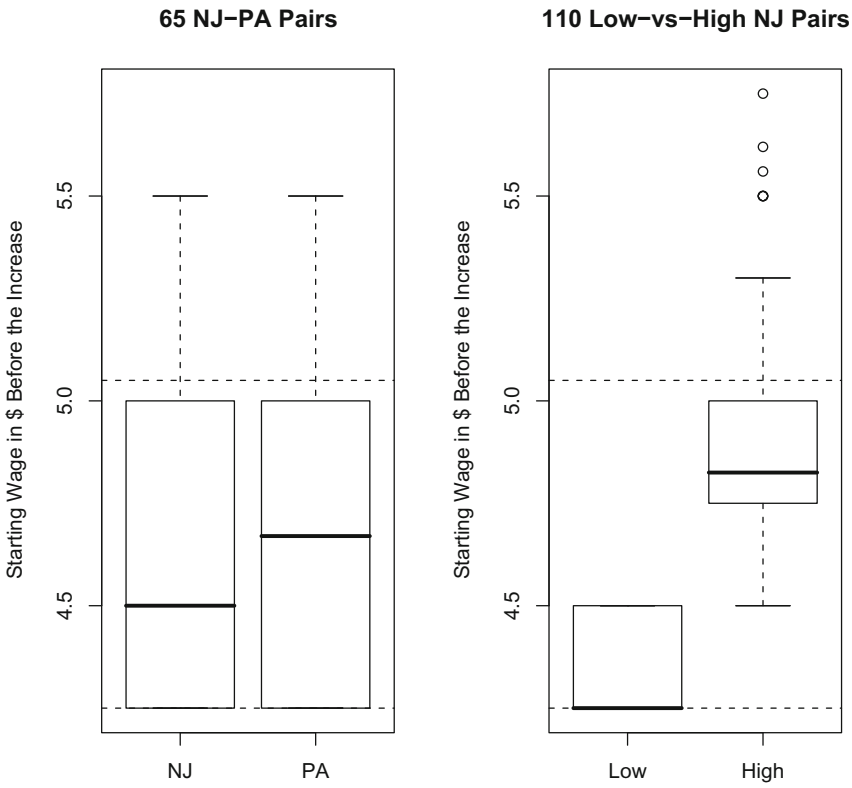

Fig. 12.1 Starting wages in dollars per hour in February 1992, before New Jersey increased its minimum wage on April 1, 1992. In the 65 pairs of a New Jersey (NJ) and a Pennsylvania (PA) restaurant, restaurants were matched to have similar starting wages, restaurant chains, company ownership, and hours open in February 1992. In the 110 pairs of two New Jersey restaurants, the matching selected one restaurant with a low starting wage and another with a high starting wage, so the low-wage restaurant was required to raise wages by a larger amount to comply with the new NJ minimum wage. The horizontal dashed lines are at the old minimum wage of $4.25 and the new minimum wage of $5.05

Table 12.9 Comparison of starting wages in New Jersey, before the increase in the minimum wage, for the 110 pairs of a low-wage (more affected) restaurant and a high-wage (less affected) restaurant

Group	Type	Mean	Min	Lower quartile	Median	Upper quartile	Max
More affected	NJ-vs.-NJ	4.33	4.25	4.25	4.25	4.5	4.50
Less affected	NJ-vs.-NJ	4.91	4.50	4.75	4.83	5.0	5.75
Difference	NJ-vs.-NJ	0.58	0.24	0.50	0.58	0.50	1.25

The minimum wage was increased to $5.05, so the less affected restaurants were, on average, forced to raise the starting wage by $0.14, while the more affected restaurants were, on average, forced to raise the starting wage by $0.72. If raising the minimum wage reduces employment, one would expect it to have a larger effect on the more affected restaurants

Table 12.10 Change in full-time-equivalent employment, after-minus-before, in All 175 pairs of a more affected and a less affected restaurant, in 65 NJ-vs.-PA pairs, and in 110 NJ-vs.-NJ pairs

Group	Type	Minimum	Lower quartile	Median	Upper quartile	Maximum
More affected	All	−20.0	−3.0	0.0	4.2	28.0
Less affected	All	−41.5	−5.0	0.0	4.0	23.5
More affected	NJ-vs.-PA	−18.5	−4.0	0.5	4.0	15.5
Less affected	NJ-vs.-PA	−41.5	−7.0	−0.5	4.0	22.8
More affected	NJ-vs.-NJ	−20.0	−2.2	0.0	4.4	28.0
Less affected	NJ-vs.-NJ	−34.0	−3.5	1.0	4.0	23.5

The NJ-vs.-PA pairs had similar starting wages before the increase in the NJ minimum wage. The NJ-vs.-NJ pairs had quite different starting wages before the increase, the less affected restaurant having a higher starting wage

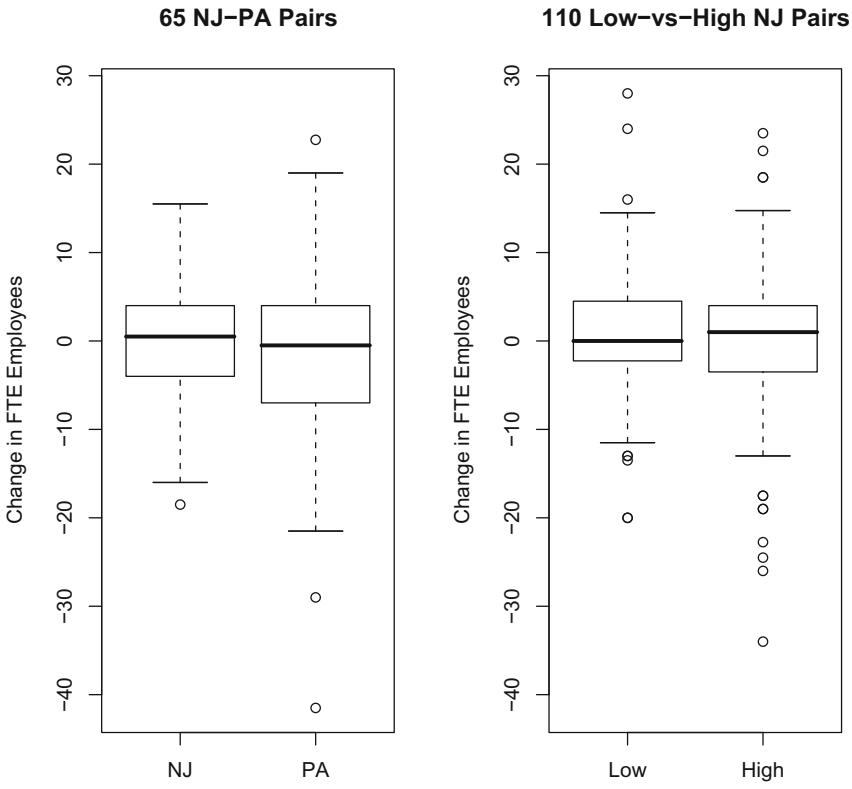

Fig. 12.2 Change in full-time-equivalent employment (FTE), after−minus−before, from before the increase in New Jersey's minimum wage to after the increase. If increasing the minimum wage tended to decrease employment at fast-food restaurants, one might reasonably expect to see larger declines in employment in the 65 New Jersey restaurants in the 65 NJ-vs.-PA pairs and in the 110 lower-wage New Jersey restaurants in the 110 NJ-vs.-NJ pairs; however, there is no sign of this

Table 12.11 Difference-in-differences estimates for FTE employment and the price of a full meal

Outcome	Type	HL-estimate	95% CI	*P*-value
Employment	All	1.00	[−0.62, 2.63]	0.22
Employment	NJ-vs.-PA	1.25	[−1.38, 4.25]	0.40
Employment	NJ-vs.-NJ	0.88	[−1.00, 2.87]	0.38
Price	All	0.040	[−0.005, 0.085]	0.07
Price	NJ-vs.-PA	0.025	[−0.045, 0.115]	0.47
Price	NJ-vs.-NJ	0.045	[−0.005, 0.105]	0.08

The matched pair difference in after-minus-before changes is examined using Wilcoxon's signed rank test and the associated Hodges–Lehmann point estimate and confidence interval. The point estimates suggest an increase, not the predicted decrease, in employment in more affected restaurants, but none of the differences in employment is significantly different from zero at the 0.05 level. There is slight evidence for an increase in the price of a full meal, but it, too, is not significantly different from zero at the 0.05 level, and the magnitude of the increase in price is estimated to be less than 5 cents

of the changes look implausibly large, whether positive or negative, perhaps due to some form of measurement error, but most values are plausible. The median change in both more and less affected restaurants was zero.

Table 12.11 compares the changes in more versus less affected restaurants. These are differences-in-differences: matched pair differences, more-minus-less affected restaurants, in the change, after-minus-before. The upper half of Table 12.11 refers to employment and the lower half to the price of a full meal, meaning soda, fries, and entree. Inferences use Wilcoxon's signed rank test, the associated confidence interval for an additive effect, and the Hodges–Lehmann point estimate of an additive effect; see Chap. 2. Although Stigler [33] predicted that increases in the minimum wage reduce employment, the point estimates for employment in Table 12.11 are positive, not negative, and none is significantly different from zero at the 0.05 level. In the "All" pairs comparison, a 1-employee decline falls outside the 95% confidence interval. There is slight evidence of a small increase in prices in more affected restaurants, but the point estimate is less than five cents, and it is not significantly different from zero at the 0.05 level.

12.4 Further Reading

Statistical uses and applications of optimal nonbipartite matching are discussed in [1, 2, 13, 14, 21–24, 27, 31, 32, 35]. As these references indicate, Bo Lu and Robert Greevy played an important role in advancing the use of nonbipartite matching in statistics; see [14, 21–24]. The specific method in Sect. 12.3 illustrates both matching with doses [23] and matching with two control groups [22]. Optimal nonbipartite matching is also useful in "risk-set" matching which is discussed in Chap. 13 and in [20, 21, 31, 32]. Nonbipartite matching permits matching

before randomization in randomized experiments [14]. It is also used to strengthen instrumental variables; see [1, 2, 13] and Sect. 5.4.

Jack Edmonds [12] developed some of the results used in algorithms for optimal nonbipartite matching. See [25, §11.3] and [8, §5.3] for textbook discussions and [7] for a survey article. Ulrich Derigs [11] provides an implementation in Fortran that Bo Lu et al. [24] has made accessible from within the statistical package `R`. An implementation in C is also available [7].

Section 12.3 uses data and ideas from the fine study by David Card and Alan Krueger [4]. An alternative analysis of their data using evidence factors is in [30]; see Chap. 20 for general discussion of evidence factors.

Nonbipartite matching is one of several methods for matching with several groups. For other methods, see Chap. 21 and [19, 26]. Similarly, nonbipartite matching is one of several techniques used in strengthening an instrument. For other techniques, see [34, Appendix I] and [35].

12.5 Software

Nonbipartite matching is implemented in the `R` package `nbpMatching` of Lu, Greevy, Xu, and Beck [24] using the algorithm of Derigs [11]. It is implemented with additional features [35] using mixed integer programming in Zubizarreta's `R` package `designmatch` package. Rigdon, Baiocchi, and Basu's [29] `R` package `nearfar` uses nonbipartite matching to strengthen instruments. Strengthening an instrument using bipartite matching is discussed by Yang et al. [34, Appendix I]. Nonbipartite matching is sometimes used in risk-set matching, as discussed in Chap. 13.

12.6 Data

David Card and Alan Krueger's [4, 5] data about minimum wages and employment are available from Princeton University's Industrial Relations Section webpage http://www.irs.princeton.edu.

Bibliography

1. Baiocchi, M., Small, D.S., Lorch, S., Rosenbaum, P.R.: Building a stronger instrument in an observational study of perinatal care for premature infants. J. Am. Stat. Assoc. **105**, 1285–1296 (2010)
2. Baiocchi, M., Small, D.S., Yang, L., Polsky, D., Groeneveld, P.W.: Near/far matching: a study design approach to instrumental variables. Health Serv. Outcomes Res. Method **12**, 237–253 (2012)

3. Cahuc, P., Carcillo, S., Zylberberg, A.: Labor Economics (2nd edn.). MIT Press, Cambridge (2014)
4. Card, D., Krueger, A.B.: Minimum wages and employment: a case study of the fast-food industry in New Jersey and Pennsylvania. Am. Econ. Rev. **84**, 772–793 (1994)
5. Card, D., Krueger, A.B.: Myth and Measurement: The New Economics of the Minimum Wage. Princeton University Press, Princeton (1995). Data: http://www.irs.princeton.edu/
6. Cochran, W.G., Cox, G.M.: Experimental Designs. Wiley, New York (1957)
7. Cook, W.J., Rohe, A.: Computing minimum-weight perfect matchings. INFORMS J. Comput. **11**, 138–148 (1999). Software: http://www2.isye.gatech.edu/~wcook/
8. Cook, W.J., Cunningham, W.H., Pulleyblank, W.R., Schrijver, A.: Combinatorial Optimization. Wiley, New York (1998)
9. Cox, D.R., Reid, N.: The Theory of the Design of Experiments. Chapman and Hall/CRC, New York (2000)
10. Daniel, S., Armstrong, K., Silber, J.H., Rosenbaum, P.R.: An algorithm for optimal tapered matching, with application to disparities in survival. J. Comput. Graph. Stat. **17**, 914–924 (2008)
11. Derigs, U.: Solving nonbipartite matching problems by shortest path techniques. Ann. Oper. Res. **13**, 225–261 (1988)
12. Edmonds, J.: Matching and a polyhedron with 0-1 vertices. J. Res. Nat. Bur. Stand. **65B**, 125–130 (1965)
13. Ertefaie, A., Small, D.S., Rosenbaum, P.R.: Quantitative evaluation of the trade-off of strengthened instruments and sample size in observational studies. J. Am. Stat. Assoc. **113**, 1122–1134 (2018)
14. Greevy, R., Lu, B., Silber, J.H., Rosenbaum, P.R.: Optimal matching before randomization. Biostatistics **5**, 263–275 (2004)
15. Hornik, R., Jacobsohn, L., Orwin, R., Piesse, A., Kalton, G.: Effects of the national youth anti-drug media campaign on youths. Am. J. Public Health **98**, 2229–2236 (2008)
16. Imai, K., van Dyk, D.A.: Causal inference with general treatment regimes: generalizing the propensity score. J. Am. Stat. Assoc. **99**, 854–866 (2004)
17. Imbens, G.W.: The role of the propensity score in estimating dose-response functions. Biometrika **87**, 706–710 (2000)
18. Joffe, M.M., Rosenbaum, P.R.: Propensity scores. Am. J. Epidemiol. **150**, 327–333 (1999)
19. Karmakar, B., Small, D.S., Rosenbaum, P.R.: Using approximation algorithms to build evidence factors and related designs for observational studies. J. Comput. Graph. Stat. **28**, 698–709 (2019)
20. Li, Y.F.P., Propert, K.J., Rosenbaum, P.R.: Balanced risk set matching. J. Am. Stat. Assoc. **96**, 870–882 (2001)
21. Lu, B.: Propensity score matching with time-dependent covariates. Biometrics **61**, 721–728 (2005)
22. Lu, B., Rosenbaum, P.R.: Optimal matching with two control groups. J. Comput. Graph. Stat. **13**, 422–434 (2004)
23. Lu, B., Zanutto, E., Hornik, R., Rosenbaum, P.R.: Matching with doses in an observational study of a media campaign against drug abuse. J. Am. Stat. Assoc. **96**, 1245–1253 (2001)
24. Lu, B., Greevy, R., Xu, X., Beck, C.: Optimal nonbipartite matching and its statistical applications. Am. Stat. **65**, 21–30 (2011)
25. Papadimitriou, C.H., Steiglitz, K.: Combinatorial Optimization: Algorithms and Complexity. Prentice Hall, Englewood Cliffs (1982)
26. Pimentel, S.D., Small, D.S., Rosenbaum, P.R.: Constructed second control groups and attenuation of unmeasured biases. J. Am. Stat. Assoc. **111**, 1157–1167 (2016)
27. Pimentel, S.D., Small, D.S., Rosenbaum, P.R.: An exact test of fit for the Gaussian linear model using optimal nonbipartite matching. Technometrics **59**, 330–337 (2017)
28. R Development Core Team.: R: A Language and Environment for Statistical Computing. R Foundation, Vienna (2007). http://www.R-project.org

29. Rigdon, J., Baiocchi, M., Basu, S.: Near-far matching in R: the nearfar package. J. Stat. Softw. **86**, 5 (2018). https://doi.org/10.18637/jss.v086.c05
30. Rosenbaum, P.R.: Evidence factors in observational studies. Biometrika **97**, 333–345 (2010)
31. Rosenbaum, P.R., Silber, J.H.: Sensitivity analysis for equivalence and difference in an observational study of neonatal intensive care units. J. Am. Stat. Assoc. **104**, 501–511 (2009)
32. Silber, J.H., Lorch, S.L., Rosenbaum, P.R., Medoff-Cooper, B., Bakewell-Sachs, S., Millman, A., Mi, L., Even-Shoshan, O., Escobar, G.E.: Additional maturity at discharge and subsequent health care costs. Health Serv. Res. **44**, 444–463 (2009)
33. Stigler, G.J.: The economics of minimum wage legislation. Am. Econ. Rev. **36**, 358–365 (1946)
34. Yang, F., Zubizarreta, J.R., Small, D.S., Lorch, S., Rosenbaum, P.R.: Dissonant conclusions when testing the validity of an instrumental variable. Am. Stat. **68**, 253–263 (2014)
35. Zubizarreta, J.R., Small, D.S., Goyal, N.K., Lorch, S., Rosenbaum, P.R.: Stronger instruments via integer programming in an observational study of late preterm birth outcomes. Ann. Appl. Stat. **7**, 25–50 (2013)

Chapter 13
Risk-Set Matching

Abstract When a treatment may be given at various times, it is important to form matched pairs or sets in which subjects are similar prior to treatment but avoid matching on events that were subsequent to treatment. This is done using risk-set matching, in which a newly treated subject at time t is matched to one or more controls who are not yet treated at time t based on covariate information describing subjects prior to time t.

13.1 Does Cardiac Transplantation Prolong Life?

In the late 1960s and early 1970s, heart transplantation was a new surgical procedure. In 1972, Mitchell Gail [9] published a brief but influential and insightful critique of two empirical studies from that time that evaluated cardiac transplantation. Hearts were not immediately available for transplantation; a candidate for a heart had to wait until one became available. In "Does cardiac transplantation prolong life? A reassessment," Gail wrote:

> Two recent reports of experience with cardiac transplantation conclude that this procedure appears to prolong life … [however] the observed differences can be explained in terms of probable selection bias. … In both [studies] the patient is assigned to the nontransplant group by default. That is, a potential transplant recipient becomes a member of the nontransplant group because no suitable donor becomes available before the potential recipient dies … [T]his assignment method biases the results in favor to the transplanted group. [9, p. 815] … [T]he survival time of the nontransplanted group is shorter than would have been observed with a random assignment method, because the method used assigns an unfair proportion of the sicker patients to this group … [and] the survival time of the transplanted group is longer than would have been observed with a random assignment method for two reasons. First, an unfairly large number of good risk patients have been assigned to this group, introducing a bias. Second, the patients in this group are guaranteed (by definition) to have survived at least until a donor was available, and this grace period has been implicitly added into the survival time of the transplanted group. [9, p. 816]

If you die early, before a heart becomes available, you are a "control." If you survive long enough for a heart to become available for you, then you become a treated subject. In an experiment, exclusion criteria are applied before random

© Springer Nature Switzerland AG 2020

P. R. Rosenbaum, *Design of Observational Studies*, Springer Series in Statistics,
https://doi.org/10.1007/978-3-030-46405-9_13

treatment assignment, ensuring that the same exclusion criteria are applied to treated and control groups (Sect. 1.2.5), but here, dying early causes you to be included in the control group and excluded from the treatment group. The term "immortal time bias" has sometimes been used to describe this situation [30]; see also [23].

The difficulty is created by the delay, waiting for a suitable heart. Gail then suggested a design for a randomized experiment that removes this difficulty. When a heart becomes available, the two living individuals most compatible with that heart are identified, paired, and then one is selected at random to receive it, the other becoming a control, with survival measured from that date of randomization [9, p. 817]. This randomized experiment addresses the problem that hearts are not immediately available, yet creates an appropriate randomized control group, one that could be used in randomization inference.

A similar but slightly different randomized experiment would begin as Gail's hypothetical experiment begins, pairing individuals when a heart becomes available, assigning the heart at random to one member of the pair; however, unlike Gail's experiment, the control, who did not receive the newly available heart, would remain eligible to receive a suitable heart if one became available at a later time.

In general, this slightly different experiment would estimate the effect of delay, that is, the effect of treating now as opposed to waiting and possibly treating later. There are contexts, common in medicine, in which this is the practical choice: treat now or wait and see; there is the option of treating later if treatment is delayed. Is treating immediately advantageous? Does it improve outcomes? Or is waiting better? Would delay lead to a better decision about who to treat? Would delay eliminate unnecessary treatment without harming outcomes? Or is treatment inevitable? If treatment is inevitable, is delay pointless or harmful? A randomized experiment of this form could answer such questions. Risk-set matching creates an observational study analogous to this randomized experiment.

13.2 Risk-Set Matching in a Study of Surgery for Interstitial Cystitis

Interstitial cystitis (IC) is a chronic urologic disorder characterized by bladder pain and irritative voiding, similar to the symptoms of urinary tract infection but without evidence of infection. In an effort to better understand the disorder and its treatment, the National Institute of Diabetes, Digestive, and Kidney Diseases created the Interstitial Cystitis Data Base (ICDB) [21]. In [13], Paul Li, Kathleen Propert, and I proposed risk-set matching as a method for studying the effects of a surgical intervention, cystoscopy, and hydrodistention, on symptoms of interstitial cystitis using data from the ICDB. Aspects of this study are described here.

To enter the ICDB, a patient must have symptoms of IC for at least the previous 6 months. Patients are evaluated upon entry into the ICDB and at approximately 3-month intervals thereafter. Three quantities were measured regularly: pain and

urgency, both recorded on scale of 0–9, where higher numbers indicate greater intensity; and nocturnal frequency of voiding. Periodically, an additional patient was treated with the surgical procedure, cystoscopy, and hydrodistention. For brevity in what follows, cystoscopy and hydrodistention are described as "surgery" or "the treatment."

Patients were not selected at random for surgery. Presumably, a patient who finds current symptoms to be intolerable is more likely to opt for surgery. This presents a problem. One cannot reasonably compare all patients in the ICDB who received surgery with all those who did not, because to know that a patient never received surgery is to have reason to suspect the patient's symptoms never became intolerable. Implicitly, to know that a patient never received surgery is to know a bit about the patient's entire course of symptoms. We would like to create pairs of patients who were similar up to the moment that one of them received surgery, but without any knowledge of their subsequent symptoms. Matching should make pairs comparable prior to treatment; what happens later is an outcome. Therefore, a new surgical patient is paired with a control patient with similar symptoms up to the point of surgery for the surgical patient; however, this control may receive surgery at a later time. This paired comparison estimates the effect of surgery now versus delaying surgery into the indefinite future, with the possibility that surgery will never be performed. It estimates the effect of the choice that patients and surgeons keep facing.

Whenever a patient received surgery, that patient was paired with a patient who had not yet received surgery but who had similar symptoms at baseline, upon entry into the ICDB, and also at the time of surgery for the surgical patient. Based on their symptoms at baseline and just prior to surgery, it would have been difficult to guess which patient would receive surgery because their symptoms at that time were similar. By definition, this control patient did not receive surgery in the 3 month recording interval following surgery for the treated patient, but the control might have surgery at any time thereafter.[1]

Figure 13.1 depicts both the baseline symptoms used for matching and the same symptoms 6 months after the date of surgery for the surgical patient. In Fig. 13.1, baseline refers to the time of entry into the ICDB, time 0 months refers to the time just prior to surgery for the surgical patient, and 3 months refers to 3 months after surgery. For instance, if a patient was treated surgically 9 months after entry into the ICDB, then for the pair that includes this patient, time 0 is 9 months after entry into the ICDB, and time 3 months is 12 months after entry into the ICDB for both patients. In Fig. 13.1, the matching appears to have been successful in the specific sense that surgical patients and their matched risk-set controls had similar distributions of symptoms at baseline and prior to the surgical patient's

[1]The matching algorithm used a distance as in Sect. 9.3, but with an important change. The distance between a patient who received surgery at time t and a patient who had not yet received surgery at time t was computed from the covariates for these two patients up to time t, without reference to information obtained after time t.

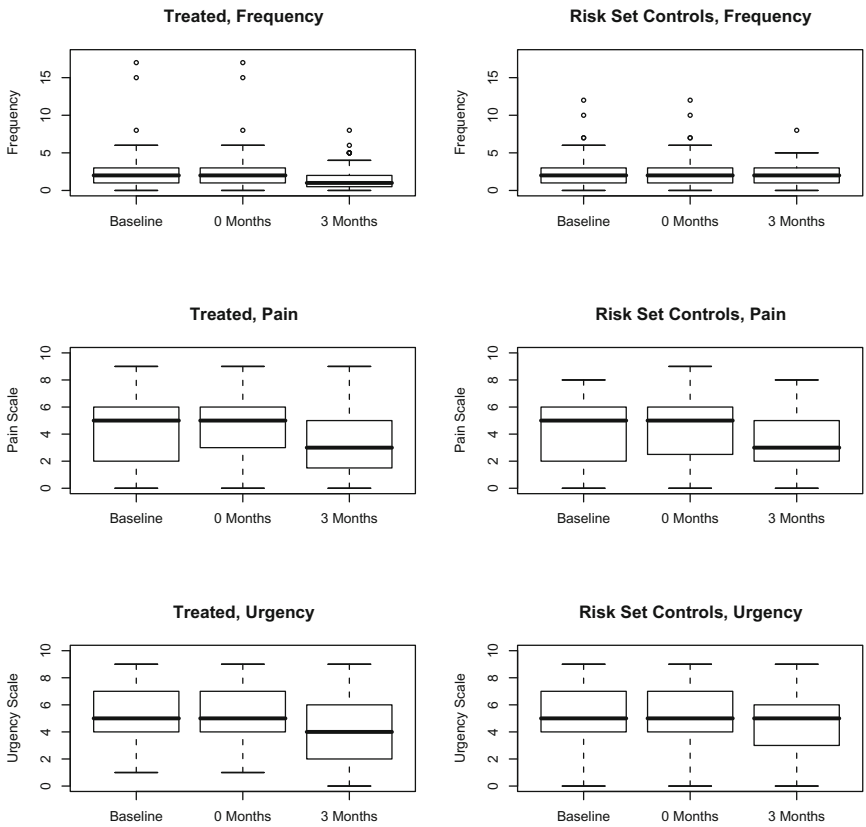

Fig. 13.1 Frequency, urgency, and pain at baseline, at treatment, and 3 months after treatment for treated patients and their matched not-yet-treated risk-set controls

date of surgery. There is a very slight uptick in the distribution of pain prior to surgery, but matching has ensured that the uptick is present for both surgical patients and matched controls. Indeed, the symptoms of the two matched individuals were similar over time. Specifically, the correlations between the symptoms of the two matched individuals were high: at baseline, for frequency, pain, and urgency, the Spearman rank correlations were, respectively, 0.92, 0.94, and 0.87, whereas at time 0 they were 0.92, 0.91, and 0.92.

In Fig. 13.1, pain and urgency scores improved 3 months after surgery for surgical patients, but they improved by just about the same amount for matched controls. In Fig. 13.1, frequency improved somewhat for surgical patients but not for controls.

Figure 13.2 sharpens this comparison. For each of the three symptoms, frequency, pain, and urgency, for each patient, the change is computed between the measurement at 3 months and the average of the measurements at baseline and at time zero. Figure 13.2 depicts the treated-minus-control matched pair difference in

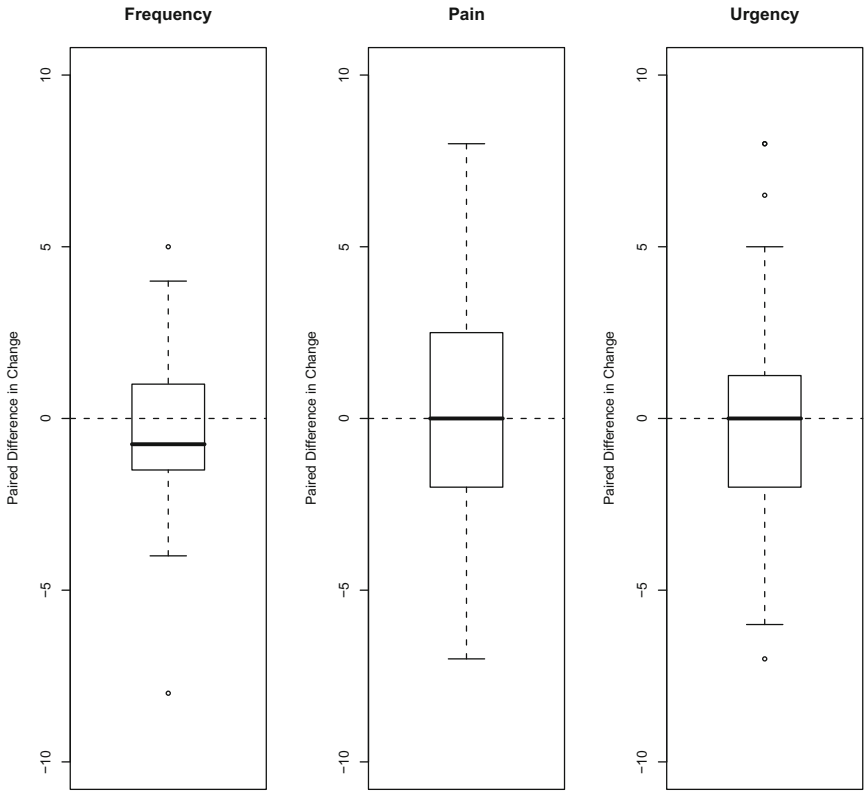

Fig. 13.2 Treated-minus-control difference in changes in frequency, urgency, and pain

Table 13.1 Inferences about the matched pair differences in changes in IC symptoms

	Frequency	Pain	Urgency
P-value	0.004	0.688	0.249
HL-estimate	−0.50	0.00	−0.50
95% CI	[−1.00, 0.00)	[−0.50, 0.75]	[−1.0, 0.25]

The two-sided P-value is from Wilcoxon's signed rank test and tests the null hypothesis that the differences are symmetric about zero. Associated with the signed rank test is the Hodges–Lehmann (HL) point estimate of the center of symmetry and the 95% confidence interval (CI) formed by inverting the test

these differences. For pain and urgency, the median difference in the changes is zero, suggesting little effect of surgery. For frequency, the median difference in the changes is −0.75, a fairly small change. Table 13.1 applies Wilcoxon's signed rank test to the matched pair differences in changes, where the difference in frequency is significantly different from zero at the 0.05 level.

13.3 Maturity at Discharge from a Neonatal Intensive Care Unit

A premature baby is kept in the neonatal intensive care unit (NICU) until it has matured sufficiently to go home [1]. Once babies have matured sufficiently to go home, they typically stay a few more days, just to be sure, and the number of days included in these "few more days" can vary considerably from one baby to the next. Recently, Jeffrey Silber, Scott Lorch, Barbara Medoff-Cooper, Susan Bakewell-Sachs, Andrea Millman, Lanyu Mi, Orit Even-Shoshan, Gabriel Escobar, and I asked [27, 29]: Does a longer stay in the NICU after maturity benefit the babies who receive it?

We looked at 1402 premature infants born in five hospitals of the Northern California Kaiser-Permanente Medical Care Program between 1998 and 2002. The relevant age for a premature baby is not age from birth but rather postmenstrual age or PMA, and gestational age refers to the age at birth. All 1402 babies in this study were born with a gestational age of 34 weeks or less and were discharged from the hospital alive. Using risk-set matching [13, 14], we divided the 1402 babies into 701 pairs of two babies, an "early baby" and a "late baby," in such a way that the two babies were similar on the day, the PMA, that the early baby was discharged, but the late baby stayed a few more days.[2] As anticipated, the late baby was not only older on its day of discharge but also heavier and had maintained various measures of maturity for a longer period of time. Did the late baby benefit from a few more days to grow older, heavier, and more mature inside the NICU? Or would the quite substantial cost of those extra days be better spent improving outpatient services for such babies?

The matching balanced the variables in Table 13.2. In Table 13.2, the first three groups of variables do not vary with time. In consequence, their values are the same in the last two columns of Table 13.2. The first group of covariates describes the babies at birth. On average, they were 31 weeks old at birth and weighed about 1.7 kg (3.75 pounds). The second group of variables describes the baby's history of significant health problems. The third group of covariates describes mom.

The fourth group of variables measures maturity, so they do vary with time; they differ in the last two columns of Table 13.2. Aspects of maturity include: (1) maintenance of body temperature, (2) coordinated sucking, (3) sustained weight gain, and (4) maturity of cardiorespiratory function. Six dimensions of maturity were scored daily as binary variables, 1 indicating that this dimension had not yet been achieved, and 0 indicating that it had been achieved. In almost all cases, babies had six zeros on their day of discharge, so the relevant question is how long the baby had a zero score before discharge. We measured this by applying exponential smoothing [6] to the binary variables, so that the smoothed scores were between 0 and 1, and a score near 0 indicated that the baby had achieved and maintained a

[2]The matching was an optimal nonbipartite matching; see Chap. 12.

Table 13.2 Balance on fixed and time-dependent covariates after risk-set matching for 1402 premature babies in 701 matched pairs

Covariate group	Covariate	Early baby at early baby discharge	Late baby at early baby discharge	Late baby at late baby discharge
	Number of babies	701	701	701
Baby at birth (fixed)	Gestational age (weeks) at birth	31.1	31.1	31.1
	Weight at birth (grams)	1669	1686	1686
	SNAP-II 20 to 59	0.15	0.13	0.13
	SNAP-II 10 to 19	0.18	0.20	0.20
	SNAP-II 0 to 9	0.67	0.67	0.67
	Male sex	0.51	0.52	0.52
Baby's health history (fixed)	Bronchopulmonary dysplasia	0.09	0.11	0.11
	Necrotizing enterocolitis	0.01	0.01	0.01
	Retinopathy stage ≥ 2	0.06	0.06	0.06
	Intraventricular hemorrhage ≥ 3	0.02	0.01	0.01
Mom (fixed)	Maternal age (years)	29.9	30.3	30.3
	Marital status single	0.24	0.24	0.24
	Other children = 0	0.40	0.37	0.37
	Other children = 1	0.34	0.37	0.37
	Other children ≥ 2	0.26	0.26	0.26
	Income $	59,517	59,460	59,460
	White race	0.47	0.48	0.48
	Black	0.10	0.09	0.09
	Asian	0.20	0.23	0.23
	Hispanic	0.22	0.18	0.18
Baby's time dependent variables	Postmenstrual age (days)	247.4	247.4	250.9
	Propensity to discharge	0.67	0.64	1.33
	Apnea smoothed score	0.04	0.05	0.03
	Brady smoothed score	0.06	0.07	0.04
	Methyl smoothed score	0.04	0.03	0.02
	Oxygen smoothed score	0.11	0.11	0.07
	Gavage smoothed score	0.22	0.23	0.10
	Incubator smoothed score	0.15	0.15	0.08
	Combined maturity score	0.62	0.63	0.34
	Current weight	2153	2148	2231
	Current weight < 1700 g	0.02	0.03	0.01
	$1700 \leq$ weight < 1800	0.06	0.06	0.02
	Current weight ≥ 1800 g	0.92	0.91	0.97

Matching ensured that paired babies were similar on the day the early baby was discharged home from the neonatal intensive care unit, but the late baby was more mature (older, heavier) on the later day of discharge for the late baby. Of course, the fixed covariates are the same on both days; only the time-dependent covariates change

Table 13.3 Absolute difference in covariate means in units of the standard deviation for 20 fixed covariates and 13 time-dependent covariates

Quantile	min	25%	50%	75%	max
Fixed	0.00	0.01	0.04	0.06	0.09
Time-dependent at early discharge	0.00	0.01	0.02	0.06	0.09
Time-dependent at own discharge	0.09	0.16	0.19	0.34	0.75

For the time-dependent covariates, two values are reported, one comparing the babies on the day the early baby was discharged, the other comparing the babies on their own days of discharge. The babies are quite similar on the day the early baby was discharged, but quite different on the day of their own discharge. On their own day of discharge, three of 13 variables had absolute standardized differences above 0.6, namely the time-dependent propensity score, the combined maturity score, and the smoothed gavage score

zero score for quite a few days. In Table 13.2, these are the Apnea, Brady, Methyl, Oxygen, Gavage, and Incubator smoothed scores; notice that these are similar on the day the early baby went home, but somewhat closer to zero for the late baby on the day the late baby went home. Another time-dependent covariate is the baby's current weight, which is recorded in several forms in Table 13.2; the late baby went home weighing about 100 grams more. Also, there is the baby's current age or PMA; the late baby was about 3.5 days older at discharge. The time-dependent propensity score [13, 14] was based on Cox's [7] proportional hazards model using both the fixed and time-dependent variables; it is the linear portion or log-hazard of discharge from the model, and it varies from day to day.

Table 13.3 summarizes Table 13.2 in terms of standardized differences; see Sect. 10.1. For the covariates in Table 13.2, absolute standardized differences in means are displayed in Table 13.3. Although the babies were similar on the day the early baby was discharged, the late baby was substantially more mature on its own day of discharge. For three time-dependent variables, the difference is more than 0.6 standard deviations.

A detailed analysis of outcomes appears in [27, 29]; see also Sects. 22.2 and 23.6. Here, it suffices to say that the early and late babies had similar experiences after discharge, so there was little indication of benefit or harm from the extra days in the NICU. Of course, the extra days were quite expensive.

The matching was implemented as follows. As noted, the proportional hazards model contributed a time-dependent propensity score [13, 14]. Then a 1402×1402 distance matrix was computed comparing the babies to one another. If the baby in row i and the baby in column j were discharged on the same day (the same PMA), then the distance between them was infinite. Otherwise, one baby was discharged earlier than the other, and the distance in row i and column j describes both babies on the day the earlier baby was discharged. The distance used a penalty to implement a caliper on the time-dependent propensity score (Sect. 9.4) and a Mahalanobis distance using the current values of the time-dependent covariates, with small penalties to improve balance on recalcitrant variables. Optimal nonbipartite matching was applied to this distance matrix to divide the 1402 babies into

701 pairs so that the total distance within pairs was minimized; see Chap. 12. The computations used Fortran code developed by Derigs [8] that Bo Lu et al. [15] made available inside `R` in the `nbpMatching` package; see also Sect. 12.5.

13.4 Joining a Gang at Age 14

Is joining a gang at age 14 a major turning point in a boy's life? Does it initiate a life of violence, a violent career? Amelia Haviland, Daniel Nagin, Richard Tremblay, and I [12] examined this question using data from the Montréal Longitudinal Study of Boys [31], which followed a cohort of boys in kindergarten in 1984 through 1995 when their average age was 17. The boys were from 53 schools in low socioeconomic areas of Montréal, Canada, and all were white, born in Canada to French-speaking parents. The data in the study are based on assessments by parents, teachers, peers, self-reports, and records from schools and the juvenile court. Violence was scored as a weighted count of violent incidents [12, p. 425]. It should be mentioned at the outset that Montréal is not one of earth's most violent places, and the boys we studied were not the most violent boys in the Montréal Longitudinal Study; see [12] for a detailed description of the group under study.

The study looked at boys who had not joined a gang prior to age 14 and who were not part of a small group of exceptional boys with very high and consistent violence prior to age 14. Among these boys, 59 joined a gang at age 14. The 59 joiners were quite different from the typical boy in this group prior to age 14: they were more violent prior to age 14, less popular, more aggressive, more oppositional, more sexually active, and their mothers were younger. Each joiner (J) was matched to two controls (C) from the same group who did not join a gang at age 14. In the 59 matched sets, the 59 joiners did not differ greatly from the $2 \times 59 = 118$ matched controls in terms of the covariates. The matching used the techniques described in Chap. 9 together with Nagin's [19] latent trajectory groups fitted to violence prior to age 14. This is the simplest form of risk-set matching because the risk-set is defined at a single time, namely age 14.

Figure 13.3 depicts the results at various ages for joiners (J) and their matched controls (C). The upper portion of the figure depicts violence, while the lower portion depicts gang membership. By definition of the study group, none of these boys were in gangs at ages 12 and 13, and all of the joiners and none of the matched controls were in gangs at age 14. Because of the matching, the boys were similar in terms of violence at ages 12 and 13. For joiners, there is a step up in violence at age 14 that is smaller at age 15 and is no longer significantly different from zero at ages 16 and 17. Most of the joiners had quit the gang by age 15, and a few of the controls had joined. By ages 16 and 17, the difference in the percentage of gang membership between joiners and controls is quite small. Only four of 59 joiners were in gangs in all 4 years, ages 14–17.

In this cohort, joining a gang at age 14 does not appear to launch a violent career: most joiners soon quit without a permanent elevation in violence, and controls often

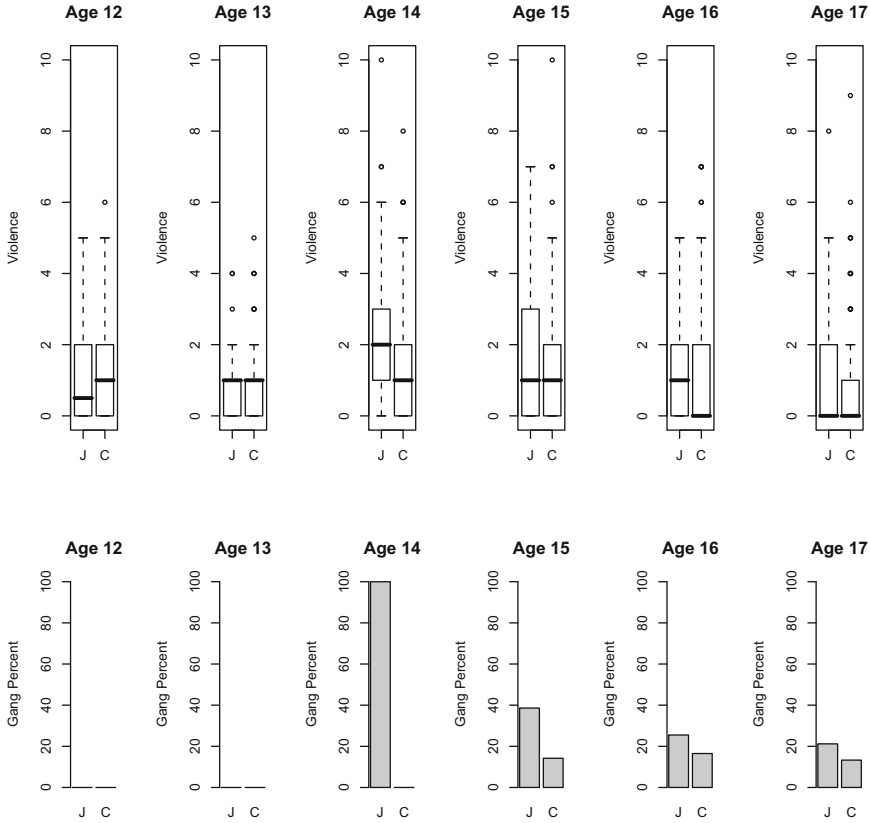

Fig. 13.3 Gang violence and gang membership for 59 boys who joined gangs for the first time at age 14 and for two risk-set controls who had not joined gangs through age 14

subsequently joined and quit gangs. Appearances suggest that joining a gang has highly transient effects. A conclusion of this sort is possible only if individuals who were similar at a certain point in time are followed forward through time; that is, the groups are not defined by what happened to individuals subsequently. One could imagine an analysis that sliced boys into boy-years, counting a boy as a gang member in years of membership and as a control in other years, but such an analysis could easily miss the highly transient nature of gang membership in the lives of intact boys.

13.5 Some Theory

The theory of risk-set matching is not difficult and is developed in about four pages in [13, §4]. A brief, informal sketch of this theory follows.

You are waiting for an event to occur, but you do not know when it will occur. In the most common medical applications, the event is death, but in risk-set matching, the event is the start of treatment, perhaps surgery, or discharge from a neonatal intensive care unit, or joining a gang. At any moment, given that the event has not yet occurred, there is a small chance it will occur in the next moment. Given that you are alive reading this sentence, there is a small chance you will not live to read its last word. These small chances, divided by the length of the moment, are called hazards.[3] If an event has not yet occurred, up to the current instant, for both of two people, and those two people have the same hazard of this event, and if the event occurs for exactly one of them, then each of the two people has an equal chance, $\frac{1}{2}$, of being the recipient of this event. This logic, expanded in various ways, underlies much of what is done with hazards, including much of Cox's proportional hazards model [7]. Suppose that two people are identical up to a certain time with regard to certain measured time-varying covariates, such as pain in Sect. 13.2, and suppose that the hazard of treatment depends only on these measured covariates, and suppose, finally, that exactly one of them is treated in the next moment; then they each have equal chance, $\frac{1}{2}$, of treatment, not unlike the "slightly different experiment" discussed at the end of Sect. 13.1. That "slightly different randomized experiment" is built in an unusual way, but the essential properties of its randomization inferences are about the same as those in Chap. 2. Under the suppositions just mentioned, exact risk-set matching for time-varying covariates recreates the randomization distribution of treatment assignments in the "slightly different randomized experiment"; in essence, it recreates the naive model of Chap. 3. The supposition that we have measured all of the covariates that affect the hazard of treatment is no small supposition; indeed, it is not especially plausible. We might imagine that there is an unmeasured time varying-covariate that affects the hazard of treatment and that was not controlled by risk-set matching for the observed covariates. With only a little structure, it is possible to show that matching for the observed time-varying covariates then reproduces the sensitivity analysis model of Chap. 3. In other words, risk-set matching for measured covariates brings us back to the two simple models of Chap. 3. Again, the technical details are not difficult and are developed in a few pages in [13, §4].

[3]This is actually a calculus notion: the moments of time grow in number while becoming shorter and shorter in duration, so the chance that anything happens in such a short interval is shrinking, but the time interval is shrinking as well, so perhaps the ratio, the hazard, tends to a limit; if so, the limit is the instantaneous hazard function at a specific time.

13.6 Isolation in Natural Experiments

13.6.1 Brief Review of Differential Effects and Generic Biases

The effect of a treatment versus no treatment may be biased by failure to adjust for an unobserved covariate relevant to the decision to treat or not treat. In Sect. 5.2.6, it was observed that, in this same situation, the effect of one treatment given in lieu of another treatment may not be biased by this same unobserved covariate. That is, there may be an unobserved generic bias that promotes several treatments at the same time, in the same way. Stated in the language of Sect. 5.2.6 and [24–26, Chapter 12], the differential effect of one treatment given in lieu of another may not be biased by a generic bias (5.2) that promotes both treatments rather than no treatment.

For instance, Sect. 5.2.6 discussed the hypothesis that perhaps NSAID medications, such as ibuprofen, reduce the risk of Alzheimer disease. Users of ibuprofen are likely to be acting to relieve pain of some sort, leading to potential biases when users are compared to nonusers. In particular, someone in the early stages of dementia may be less aware of pain or less able to act effectively in response to that awareness. However, this is a generic bias: it affects the use of pain relievers generally. Acetaminophen is a pain reliever that is not an NSAID. It is possible that the main effect of ibuprofen versus no medication is biased, but the comparison of ibuprofen versus acetaminophen is not biased, yet both comparisons compare an NSAID with the absence of an NSAID [3].

The issue just described has a temporal dimension that was not considered in Sect. 5.2.6. Over time, an individual is likely to use a pain reliever when in pain, and to abstain from use when not in pain. Over time, the frequency and timing of the use of ibuprofen rather than no medication may indicate unmeasured fluctuating levels of pain, levels of awareness of pain, and levels of activity in response to awareness. However, given that an individual used a pain reliever at a given moment in time, the decision to use ibuprofen rather than acetaminophen may not be biased by these unmeasured fluctuating covariates. This leads to the notion of isolation introduced by José Zubizarreta et al. [36]: differential effects assessed in a risk-set match.

13.6.2 What Is Isolation?

Isolation refers to examination of the differential effect of two treatments given a moment in time when one of those two treatments was received in lieu of the other. It compares two matched people who looked similar up to that moment, whereupon one person received the first treatment and the other person received the second treatment. Isolation imagines a situation in which the occurrence of treatment, the frequency of treatment, the timing of treatment are all biased by unmeasured covariates; however, given that two similar people received different treatments at a

certain moment, there is no bias in who received which treatment.[4] Isolation creates matched pairs: two people who were similar up to the moment when they both received treatment, albeit different treatments.

13.6.3 Twins Versus a Single Birth, and Their Impact on Labor Supply

To what extent do women withdraw from the workforce in response to having a child? For many women, plans for fertility and career are two aspects of one plan, not a career outcome of a fertility decision. If we saw that women with many children had limited participation in the workforce, then we would not know whether having a child interrupts a woman's career, or whether a limited career and a large family were both aspects of her plan from the beginning. And yet, there are aspects of fertility that are unplanned, so it may be reasonable to ask about the effects on a career of an unplanned aspect of fertility.

An interesting study by Joshua Angrist and William Evans [2] compared women who had had twins to women who had had a single child. In the time period covered by their data, fertility treatment was less common than it is today, and having twins was a largely unpredictable event. Do women who have twins go on to work less? If so, how much less?

Using the data from Angrist and Evans, Zubizarreta et al. [36] reanalyzed the situation as an isolated comparison. The timing of fertility was viewed as partly under a woman's control, reflective of her unmeasured plans for fertility and career. Given that a woman is pregnant at a certain moment, and given covariates that describe her fertility, education, and career prior to that moment, having twins rather than a single child may be little more than luck. To keep the story short and to focus on the concept of isolation, the description that follows is simpler than the analysis in [2] and the reanalysis in [36].

Beginning with the second pregnancy, Zubizaretta et al. [36] matched each birth resulting in twins to five births resulting in a single child, 3380 matched sets of size 1-to-5. These 6×3380 women were followed forward in time, but were not

[4]In somewhat more technical terms, isolation refers to a marked point process extending over time. The point process determines the timing of treatment. The marks indicate the specific treatment given at each time when some treatment was given. Isolation imagines that the point process has a hazard that is affected by unmeasured covariates, so the timing of treatment is biased. However, given that a treatment is applied at a certain moment, the marks, the specific treatment at that time, is not biased by these unmeasured covariates. More generally, both points and marks may be biased, but generic unmeasured biases that affect the points but not the marks are removed by risk-set matching of a differential comparison, leaving behind only differential unmeasured biases that continue to affect the marks [36, §2.2]. In effect, this produces an ordinary matched pairs comparison of the differential effect of two treatments with the unmeasured generic biases removed.

eligible to be matched at the third pregnancy. At the third pregnancy, each birth resulting in twins was matched to 5 births resulting in a single child, yielding 1358 additional 1-to-5 matched sets. The process was repeated at the fourth pregnancy, yielding an additional 302 sets matched 1-to-5. The matching controlled for the past, not the future. Covariates described a woman prior to her current pregnancy, so covariates fluctuated with time. Again, the timing of pregnancies is thought to be biased by a woman's unmeasured plans for career and family, but given a pregnancy and similar past covariates, the occurrence of twins rather than a single child is thought to be, if not random, at least unaffected by large unmeasured biases. In other words, Zubizaretta et al. [36] isolated the almost random element, twins or single child, from highly nonrandom aspects of fertility, creating a natural experiment with $3380 + 1358 + 302$ matched 1-to-5 sets.

It appeared that having a twin in the second pregnancy or later typically, though not always, increased the ultimate family size, with the median increasing by one child, but it appeared to only slightly decrease workforce participation [36, Figure 3]. Although the reduction in participation was estimated to be small, it was insensitive to small biases in who has a twin ($\Gamma = 1.2$ in [36, Table 3]). Angrist and Evans [2] had reached qualitatively similar conclusions from a different methodological perspective.

13.6.4 Mass and Safety in Vehicle Crashes Resulting in a Fatality

It is possible to consistently drive a Camero at well below the speed limit, and to burn rubber in a Volvo when the traffic light turns green, but these possibilities defy our everyday expectations. When we see a muscle car, we do not expect to see it driven in an exceedingly cautious way. When we see a car loaded with expensive safety equipment, we do not expect to see it driven in an exceedingly reckless way. There are, without doubt, exceptions; however, these exceptions are exceptions. When studying the effects of attributes of cars on the risk of death in a vehicle crash, we have reason to worry that attributes of cars are related to the ways cars are driven. There is reason to worry that cars with expensive safety equipment are driven more safely, and so are involved in fewer and less severe crashes. One needs to avoid attributing to safety equipment effects that are actually produced by safer driving.

Engineers evaluate crash protection by crashing vehicles into immovable walls. This removes variation in the way vehicles are driven. Do most accidents involve immovable walls? In truth, crashing into an immovable wall is an uncommon way to die. Moreover, these studies remove an aspect of a vehicle that is relevant to safety, namely its mass as compared to the mass of the vehicle with which it collides. How important is mass?

Using the Fatality Analysis Reporting System (FARS), Zubizarreta and colleagues [37] looked at crashes with a fatality in which precisely two vehicles collided. Volvos may be driven more safely than Cameros, but this is less relevant given that a Volvo and a Camero have collided. Comparing two vehicles in a single crash equates some things that might otherwise have been quite different: there was a crash, in certain weather and road conditions, with certain forces that need to dissipate to bring the vehicles to a halt. Safer driving may bias the number of crashes and their severity, but all that matters less given that two vehicles have collided. Given a fatal collision of two vehicles, how much does mass matter? What is the differential effect on survival of being in the less massive vehicle in a two vehicle crash, conditionally given such a crash? What is the isolated effect?

Zubizarreta et al. [37] built one 2×2 table for each of 2923 fatal crashes involving precisely one occupied car and one occupied light truck. Light trucks included SUVs, minivans, and pick-up trucks. Each 2×2 table recorded as rows car-or-light truck, and as columns survived-or-died. By definition, every 2×2 table described two occupied vehicles, so the total count in each row was at least one. Because the Fatality Analysis Reporting System only provides information about crashes with a fatality, every 2×2 table included at least one death, so the column for death had a count of at least one. If at least one person survived the crash, then the total count in the column for survivors was at least one, and in this case the table is informative about who lives and who dies. What does informative mean here?

A standard test for partial association in a $2 \times 2 \times k$ contingency table is the Mantel–Haenszel–Birch test [5, 16].[5] This test is unaffected by a 2×2 subtable with a zero count in a marginal total.[6] Because such a table makes a degenerate or constant contribution to the test, we say that such a subtable is noninformative. Stated very informally, if everyone dies in a crash, the Mantel–Haenszel–Birch test writes that subtable off as a particularly severe crash, uninformative about the relative safety of two vehicles. In parallel, if everyone survives in a crash, the Mantel–Haenszel–Birch test writes that subtable off as a particularly mild crash, uninformative about the relative safety of two vehicles.

There are an enormous number of crashes that are not recorded in FARS because everyone involved in the crash survived. Is this a problem? Is it a problem that we do not have access to the 2×2 tables for crashes in which everyone survived? No doubt, those tables would be interesting for many purposes, but they are uninformative for the Mantel–Haenszel–Birch test: if we had those tables, the test would ignore them anyway.

[5] Birch [5] developed the exact test that corresponds with the earlier large sample test proposed by Mantel and Haenszel [16]. The null distribution of Birch's exact test is a randomization distribution. Birch also showed that the exact test is a uniformly most powerful unbiased test against certain alternative hypotheses. The `mh` function in the `sensitivity2x2xk` package in `R` performs a sensitivity analysis for the exact Mantel–Haenszel–Birch test; see [28].

[6] More precisely, such a subtable adds a constant, not a random variable, to the overall test statistic, so that its P-value is unchanged.

In the $2 \times 2 \times 2923$ table, the odds of dying in the car were estimated to be 4.76 times greater than in the light truck, with 95% confidence interval [4.34, 5.23]. If an adjustment is made for safety belt use, the odds ratio increases from 4.76 to 5.35.

Unfortunately, car versus light truck is not quite the right comparison. Some light trucks are not massive: the Chrysler PT Cruiser is nominally a "light truck," but it weighs somewhat less than a Honda Accord, a car that is not especially massive. Some cars from Lincoln and Mercedes Benz weigh as much as a typical SUV. So the $2 \times 2 \times 2923$ was split into two parts. In the first part with 1565 2×2 tables, the car was not exceptionally heavy and the light truck was not exceptionally light; specifically, the car weighed at most the upper quartile of car weights, 3442 pounds, and the light truck weighed at least the lower quartile of light truck weights, 3847 pounds. The second part contained the 1358 remaining 2×2 tables, for instance, heavy cars colliding with very light light-trucks. In the first part, the odds of death in the car were estimated to be 8.10 times greater than in the light truck, with 95% confidence interval [6.99, 9.39]. If an adjustment is made for safety belt use, the odds ratio increases from 8.10 to 9.16. In the second part of the table, the odds ratio was much smaller, 2.91 with 95% confidence interval [2.57, 3.30].

It appears that relative mass matters a great deal for survival when two vehicles collide. Although comparing two vehicles with different masses in the same crash removes many generic biases, the comparison is not immune to unmeasured bias. Conceivably, if not entirely plausibly, perhaps frail people drive lighter cars. Nonetheless, the differential effect described above is insensitive to large unmeasured biases [37].

13.7 Further Reading

Cardiac transplantation has advanced since the 1970s, but Mitchell Gail's [9] short paper is of continuing methodological interest. Risk-set matching refers to matching from the set of people who are at risk of receiving the treatment. The theory of risk-set matching is related to hazard models, such as David Cox's [7] proportional hazards model, where a time-varying hazard of treatment acts in a manner similar to the propensity score; the theory is discussed in [13, §4]. Bo Lu [14] develops the balancing properties of time-dependent propensity scores, and studies their use in conjunction with nonbipartite matching. Sue Marcus et al. [17] discuss balancing covariates in longitudinal studies. In risk-set matching, sensitivity of conclusions to bias from unmeasured time-varying covariates is discussed in [13]. The examples in Sects. 13.2–13.4 are discussed in greater detail in [12, 13, 27, 29]. An alternative approach to matching in the Montréal Longitudinal Study of Boys is discussed in [11], where many more controls are used by matching a variable number of controls to each joiner; see also Sect. 9.5. Paul Nieuwbeerta et al. [20] used risk-set matching in a study of the possible effects of first imprisonment on the development of a criminal career. Wei Wu et al. [35] study the effects of retention in first grade by matching retained students to students who appeared to be similar at the end of first

grade but were promoted to second grade, following them forward for several years. Jamie Robins and colleagues [22] have developed a general approach to treatments that evolve over time; see the book by Mark van der Laan and Jamie Robins [32] for detailed discussion of this approach.

An elementary discussion of isolation is given in [37], and the formal theory is in [36, §2.3]. Isolation is also discussed in [26, Chapter 12].

13.8 Software

Sometimes "time" is measured in coarse units: whole years, or quarters of a year, or months, or as second pregnancy, third pregnancy, fourth pregnancy. In these situations, risk-set matching can be implemented as a sequence of conventional matches. Match year one; remove matched individuals; match year two; remove matched individual; and so on. In these situations, standard software is used several times. This is the simplest approach, perhaps the best approach to risk-set matching.

An alternative approach does one match using nonbipartite matching, the topic of Chap. 12. This approach can be used with a continuous measure of "time." Fortran code developed by Derigs [8] for nonbipartite matching was made available in `R` by Bo Lu and colleagues [15] in the `nbpMatching` package; see also Sect. 12.5. The match in Sect. 13.3 used optimal nonbipartite matching.

Bibliography

1. American Academy of Pediatrics.: Hospital discharge of the high-risk neonate – proposed guidelines. Pediatrics **102**, 411–417 (1998)
2. Angrist, J.D., Evans, W.N.: Children and their parents' labor supply: evidence from exogenous variation in family size. Am. Econ. Rev. **88**, 450–477 (1996)
3. Anthony, J.C., Breitner, J.C., Zandi, P.P., Meyer, M.R., Jurasova, I., Norton, M.C., Stone, S.V.: Reduced prevalence of AD in users of NSAIDs and H_2 receptor antagonists. Neurology **54**, 2066–2071 (2000)
4. Apel, R., Blokland, A.A.J., Nieuwbeerta, P., van Schelllen, M.: The impact of imprisonment on marriage and divorce: a risk-set matching approach. J. Quant. Criminol. **26**, 269–300 (2010)
5. Birch, M.W.: The detection of partial association, I: the 2 x 2 case. J. R. Stat. Soc. B **26**, 313–324 (1964)
6. Cox, D.R.: Prediction by exponentially weighted moving averages and related methods. J. R. Stat. Soc. B **23**, 414–422 (1961)
7. Cox, D.R.: Regression models and life-tables. J. R. Stat. Soc. B **34**, 187–220 (1972)
8. Derigs, U.: Solving nonbipartite matching problems by shortest path techniques. Ann. Oper. Res. **13**, 225–261 (1988)
9. Gail, M.H.: Does cardiac transplantation prolong life? A reassessment. Ann. Intern. Med. **76**, 815–817 (1972)
10. Haviland, A.M., Nagin, D.S.: Causal inferences with group based trajectory models. Psychometrika **70**, 557–578 (2005)

11. Haviland, A.M., Nagin, D.S., Rosenbaum, P.R.: Combining propensity score matching and group-based trajectory analysis in an observational study. Psychol. Methods **12**, 247–267 (2007)
12. Haviland, A.M., Nagin, D.S., Rosenbaum, P.R., Tremblay, R.E.: Combining group-based trajectory modeling and propensity score matching for causal inferences in nonexperimental longitudinal data. Dev. Psychol. **44**, 422–436 (2008)
13. Li, Y.F.P., Propert, K.J., Rosenbaum, P.R.: Balanced risk set matching. J. Am. Stat. Assoc. **96**, 870–882 (2001)
14. Lu, B.: Propensity score matching with time-dependent covariates. Biometrics **61**, 721–728 (2005)
15. Lu, B., Greevy, R., Xu, X., Beck, C.: Optimal nonbipartite matching and its statistical applications. Am. Stat. **65**, 21–30 (2011)
16. Mantel, N., Haenszel, W.: Statistical aspects of the analysis of data from retrospective studies of disease. J. Nat. Cancer Inst. **22**, 719–748 (1959)
17. Marcus, S.M., Siddique, J., Ten Have, T.R., Gibbons, R.D., Stuart, E., Normand, S.-L.: Balancing treatment comparisons in longitudinal studies. Psychiatric Ann. **38**, 805–812 (2008)
18. Messmer, B.J., Nora, J.J., Leachman, R.D., et al.: Survival-times after cardiac allografts. Lancet **1**, 954–956 (1969)
19. Nagin, D.S.: Group-Based Modeling of Development. Harvard University Press, Cambridge (2005)
20. Nieuwbeerta, P., Nagin, D.S., Blokland, A.A.J.: The relationship between first imprisonment and criminal career development: a matched samples comparison. J. Quant. Criminol. **25**, 227–257 (2009)
21. Propert, K.J., Schaeffer, A.J., Brensinger, C.M., Kusek, J.W., Nyberg, L.M., Landis, J.R.: A prospective study of interstitial cystitis: results of longitudinal followup of the interstitial cystitis data base cohort. J. Urol. **163**, 1434–1439 (2000)
22. Robins, J.M., Blevins, D., Ritter, G., Wulfsohn, M.: G-estimation of the effect of prophylaxis therapy for pneumocystis carinii pneumonia on the survival of AIDS patients. Epidemiology **3**, 319–336 (1992)
23. Rosenbaum, P.R.: The consequences of adjustment for a concomitant variable that has been affected by the treatment. J. R. Stat. Soc. A **147**, 656–666 (1984)
24. Rosenbaum, P.R.: Differential effects and generic biases in observational studies. Biometrika **93**, 573–586 (2006)
25. Rosenbaum, P.R.: Using differential comparisons in observational studies. Chance **26**(3), 18–23 (2013)
26. Rosenbaum, P.R.: Observation and Experiment: An Introduction to Causal Inference. Harvard University Press, Cambridge (2017)
27. Rosenbaum, P.R., Silber, J.H.: Sensitivity analysis for equivalence and difference in an observational study of neonatal intensive care units. J. Am. Stat. Assoc. **104**, 501–511 (2009)
28. Rosenbaum, P.R., Small, D.S.: An adaptive Mantel–Haenszel test for sensitivity analysis in observational studies. Biometrics **73**, 422–430 (2017)
29. Silber, J.H., Lorch, S.L., Rosenbaum, P.R., Medoff-Cooper, B., Bakewell-Sachs, S., Millman, A., Mi, L., Even-Shoshan, O., Escobar, G.E.: Additional maturity at discharge and subsequent health care costs. Health Serv. Res. **44**, 444–463 (2009)
30. Suissa, S.: Immortal time bias in pharmacoepidemiology. Am. J. Epidemiol. **167**, 492–499 (2008)
31. Tremblay, R.E., Desmarais-Gervais, L., Gagnon, C., Charlebois, P.: The preschool behavior questionnaire: stability of its factor structure between culture, sexes, ages, and socioeconomic classes. Int. J. Behav. Dev. **10**, 467–484 (1987)
32. van der Laan, M., Robins, J.: Unified Methods for Censored Longitudinal Data and Causality. Springer, New York (2003)
33. Watson, D., Spaulding, A.B., Dreyfus, J.: Risk-set matching to assess the impact of hospital-acquired bloodstream infections. Am. J. Epidemiol. **188**, 461–466 (2019)

34. Wermink, H., Blokland, A., Nieuwbeerta, P., Nagin, D., Tollenaar, N.: Comparing the effects of community service and short-term imprisonment on recidivism: a matched samples approach. J. Exp. Criminol., **6**, 325–349 (2010)
35. Wu, W., West, S.G., Hughes, J.N.: Effect of retention in first grade on children's achievement trajectories over 4 years: a piecewise growth analysis using propensity score matching. J. Educ. Psychol. **100**, 727–740 (2008)
36. Zubizarreta, J.R., Small, D.S., Rosenbaum, P.R.: Isolation in the construction of natural experiments. Ann. Appl. Stat. **8**, 2096–2121 (2014)
37. Zubizarreta, J.R., Small, D.S., Rosenbaum, P.R.: A simple example of isolation in building a natural experiment. Chance **31**, 16–23 (2018)

Chapter 14
Matching in R

Abstract Simple calculations in the statistical language R illustrate the computations involved in one simple form of multivariate matching. The focus is on how matching is done, not on the many aspects of the design of an observational study. The process is made tangible by describing it in detail, step-by-step, closely inspecting intermediate results; however, essentially, three steps are illustrated: (1) creating a distance matrix, (2) adding a propensity score caliper to the distance matrix, and (3) finding an optimal match. In practice, matching involves bookkeeping and efficient use of computer memory that are best handled by dedicated software for matching. Sect. 14.10 describes currently available software.

14.1 Optimal Matching Using R

Optimal multivariate matching is easy to do in the statistical package R, as is illustrated in a simple example in the current chapter. It is convenient to use one data set to illustrate various types of matching. For that reason, and for that reason only, four matched samples will be constructed from Susan Dynarski's [15] fine study of the termination in 1982 of the Social Security Student Benefit Program, which provided tuition assistance to the children of deceased Social Security beneficiaries; this study was discussed in Sect. 1.5. The construction of one of the four matched samples is presented in step-by-step detail. Outcomes are briefly compared in Sect. 14.8, but the focus is on constructing matched control groups in R. The chapter is structured to illustrate various aspects of R, not to produce a recommended analysis for Dynarski's [15] study.

In the R package DOS2, the example in the help-file for dynarski will reproduce many of the calculations in this chapter. Section 14.10 recommends various R packages that implement additional matching techniques, and these packages contain additional examples. See also Sect. 14.11.

Optimal matching is easy in R, largely due to the efforts of Ben Hansen [17], Bo Lu, Robert Greevy, Xinyi Xu, Cole Beck [23], José Zubizarreta and Cinar Kilcioglu [21, 48], Mike Baiocchi and colleagues [3, 4, 30], Samuel Pimentel [27, 28], and

© Springer Nature Switzerland AG 2020

P. R. Rosenbaum, *Design of Observational Studies*, Springer Series in Statistics,
https://doi.org/10.1007/978-3-030-46405-9_14

Ruoqi Yu [46, 47], who created R packages for that task. These packages use code from Demetri Bertsekas and Paul Tseng [9], Ulrich Derigs [14], gurobi or Cplex.

14.2 Data

The first match, the only one considered in detail, concerns seniors in the period 1979–1981, when the Social Security Student Benefit Program was still in operation; see the left half of Fig. 1.2. During this period, a child of a deceased Social Security beneficiary might receive a substantial tuition benefit from the program.

Eight covariates are used in the match.[1] They are: family income (faminc), income missing (incmiss), black, Hispanic, the Armed Forces Qualifications Test (afqtpct), mother's education (edm), mother's education missing (edmissm), and gender (female);[2] these are contained in Xb with 2820 rows and eight columns. The treatment indicator zb is a vector with a one for a senior whose father is deceased and 0 for other seniors. The first 20 of 2820 rows of the treatment indicator, zb and covariates Xb are displayed in Table 14.1. In R, a variable such as faminc in a data.frame such as Xb is referred to as Xb$faminc.

Mother's education but not father's education was used because one group was defined by a deceased father. The AFQT is the only educational measure among the covariates. The AFQT was missing for less than 2% of subjects, and these subjects are not used in the matching. With this restriction,[3] there are 131 high school seniors with deceased fathers and 2689 other high school seniors in the 1979–1981 cohort,

[1] Dynarski [15, Table 2] presents two analyses, one with no covariates, the other with many more covariates, obtaining similar estimates of effect from both analyses. That is a reasonable approach in the context of her paper. In constructing a matched control group in this chapter, I have omitted a few of the covariates that Dynarski used, including "single-parent household" and "father attended college," from a comparison involving a group defined by a deceased father. In general, if one wants to present parallel analyses with and without adjustment for a particular covariate, say "single-parent household," then one should not match on that covariate, but should control for it in one of the parallel analyses using analytical techniques; e.g., Sect. 22.2 or [16, 32, 37, 42] and [36, §3.6].

[2] In detail, the covariates used in the match are (1) faminc: family income in units of $10,000; (2) incmiss: income missing (incmiss=1 if family income is missing, incmiss=0 otherwise); (3) black (black=1 if black, black=0 otherwise), (4) hispanic (hispanic=1 if hispanic, hispanic=0 otherwise), (5) afqtpct: Armed Forces Qualifications Test (AFQT), (6) edmissm: mother's education missing (edmissm=1 if missing, edmissm=0 otherwise), (vii) edm: mother's education (edm=1 for less than high school, edm=2 for high school, edm=3 for some college, edm=4 for BA degree or more), (viii) female (female=1 for female, female=0 for male).

[3] Because of this restriction, the counts of seniors in various groups are slightly different than in [15].

Table 14.1 For the first 20 people in the 1979–1981 cohort, the treatment `zb` and the eight covariates `Xb` are listed

id	zb	faminc	incmiss	black	hisp	afqtpct	edmissm	edm	female
1	0	5.3	0	0	0	71.9	0	2	0
2	0	2.5	0	0	0	95.2	0	2	1
4	0	9.5	0	0	0	95.8	0	2	1
5	0	0.4	0	0	0	90.6	0	2	1
6	0	10.4	0	0	0	81.8	0	2	0
7	0	6.3	0	0	0	97.9	0	4	0
10	1	3.2	0	0	0	61.9	0	2	1
11	0	9.5	0	0	0	20.4	0	2	1
12	0	2.7	0	0	0	57.2	0	1	0
14	0	2.0	1	0	0	8.3	0	1	1
15	0	7.1	0	0	0	50.9	0	2	1
16	0	7.8	0	0	0	71.1	0	2	1
17	0	12.5	0	0	0	74.9	0	2	0
18	0	2.0	1	0	0	98.4	0	2	1
21	0	9.4	0	0	0	98.7	0	4	1
23	0	2.0	1	0	0	99.7	0	4	0
24	0	2.0	1	0	0	73.3	0	3	1
29	0	2.0	1	1	0	1.8	0	1	0
31	0	2.0	1	0	0	42.9	0	2	1
34	1	2.6	0	0	0	99.6	0	2	0

Here, `zb` is 1 if the father is deceased, and `zb` is 0 otherwise. In `R`, this table is displayed by `cbind(zb,Xb)[1:20,]`, which instructs `R` to bind the vector `zb` and the matrix `Xb` as columns, and to display rows 1–20 and all columns. The id numbers (actually `rownames` in `R`) are not consecutive because `zb` and `Xb` contain only the 1979–1981 cohort, before the cancellation of the Social Security Student Benefit Program in 1982. As discussed in Chap. 10, when missing value indicators are used in matching (here `incmiss` and `edmissm`), the missing values in the variables themselves are arbitrary. Here, `faminc` is set to 2 when `incmiss` is 1 indicating that `faminc` is missing, and `edm` is set to 0 when `edmissm` is 1 indicating that `edm` is missing. Had different values been filled in, the propensity scores and conventional Mahalanobis distances would have been the same

before the Social Security Student Benefit Program ended in 1982; see Table 14.2. Also, $131 + 2689 = 2820$. In consequence, R finds:

```
> length(zb)
  [1] 2820
> dim(Xb)
  [1] 2820 8
> sum(zb)
  [1] 131
> sum(1-zb)
  [1] 2689
```

Table 14.2 Frequency table showing the number of seniors available and the number used in matching

	1979–1981	1982–1983	Matched 1979–1981
Father Deceased (FD)	131	54	131
Father Not Deceased (FND)	2689	1038	1038
Matched FND (10 controls)	1310	540	

In 1979–1981 and 1982–1983, seniors whose fathers were deceased are matched to ten seniors whose fathers were not deceased, yielding, respectively, 131 and 54 matched sets of 11 seniors each. Also, 1038 pairs of seniors whose fathers are not deceased are found, with one from 1979–1981 and the other from 1982–1983. Among seniors with deceased fathers, a variable match is also found, using all 54+131 seniors, with 54 matched sets containing one senior in 1982–1983 and between 1 and 4 seniors in 1979–1981

The 131 seniors with deceased fathers will each be matched to 10 controls whose fathers were not deceased.

14.3 Propensity Score

The propensity score is estimated using a logit model. In R, logit models are fitted using the metaphor of generalized linear models [25]; however, if logit models are familiar and generalized linear models are unfamiliar, then you may think of this as R's syntax for logit regression because it is the same logit regression. In R, the estimated propensity scores are the fitted probabilities of a deceased father given the eight covariates, and one obtains these as the `$fitted.values` in a generalized linear model `glm` with the `family=binomial`. Specifically, the vector `p` contains the 2820 estimated propensity scores, $\widehat{e}(\mathbf{x}_\ell)$, $\ell = 1, 2, \ldots, 2820$:

```
> p<-glm(zb~Xb$faminc+Xb$incmiss+Xb$black+Xb$hisp
  +Xb$afqtpct+Xb$edmissm+Xb$edm+Xb$female,
    family=binomial)$fitted.values
```

It is reasonable to work to improve this model, perhaps including interaction terms or transformations or polynomials or whatnot, but this chapter will take a rather minimal approach, seeking an acceptable match in a few simple steps.

Figure 14.1 displays the estimated propensity scores for the 131 seniors with deceased fathers and the other 2689 high school seniors. The difference between the two distributions is substantial: (1) the standardized difference in means (Sect. 10.1) of $\widehat{e}(\mathbf{x}_\ell)$ is 0.67, or $\frac{2}{3}$ of a standard deviation, and (2) in Fig. 14.1, the median $\widehat{e}(\mathbf{x}_\ell)$ in the treated group is about equal to the upper quartile among potential controls. Nonetheless, the distributions in Fig. 14.1 overlap substantially, so matching appears to be possible.

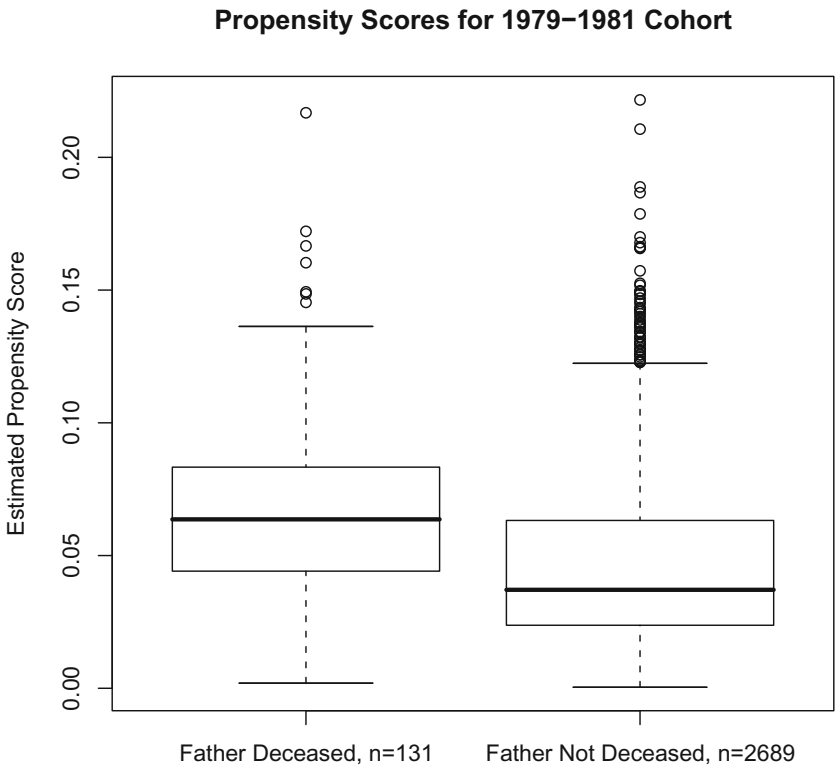

Fig. 14.1 Estimated propensity scores for 131 seniors with deceased fathers and 2689 other seniors in the 1979–1981 cohort, before the Social Security Student Benefit Program was eliminated

14.4 Covariates with Missing Values

Missing covariate values are discussed in Sect. 10.6; please review that discussion. The situation with missing covariate values is a common source of misunderstanding. As in Sect. 10.6, missing covariate values are filled in with an arbitrary but fixed number; then, missing value indicators are appended as additional covariates; see Table 14.1 and the logit model for the propensity score in Sect. 14.3. If the missing values for family income, `faminc`, or mother's education, `edm`, had been filled in with different constants, the estimated propensity scores $\widehat{e}(\mathbf{x}_\ell)$ would have been the same because of the presence of the missing value indicators, `incmiss` and `edmissm`. The coefficients in the logit model would change if different values had been used, but they change to compensate for the change in the filled-in values,

keeping $\widehat{e}(\mathbf{x}_\ell)$ unchanged.[4] To prevent round-off error in the computer's arithmetic work, it makes sense to fill in values that are not wild—a family income of -9999 is not recommended—but aside from round-off error, the value filled in does not matter. To emphasize, the missing values are still missing. The propensity score is estimating the probability that a senior will have a deceased father when that senior has certain specific measured covariates and a specific pattern of missing covariates; it is not guessing what the missing covariate values might be. That is, the propensity score is estimating a particular conditional probability about observable events, because the patterns of missing data (but not the missing values themselves) are observable. Matching or stratifying on this particular conditional probability relating observables tends to balance those observables, that is, the observed covariates and the observed pattern of missing data, but there is no basis for expecting it to balance the missing values [40, Appendix]. Specifically, a successful match should produce a similar distribution of observed family incomes (`faminc`, when `incmiss`=0) in treated and control groups and a similar percent of seniors with missing family incomes (`incmiss`=1); however, whether or not the missing incomes themselves (`faminc`, when `incmiss`=1) are similar is not known. With regard to missing covariates, the propensity score succeeds in its ambitions only because its ambitions are quite limited, namely balancing observable quantities.

14.5 Distance Matrix

The distance matrix is constructed in two steps. The first step computes the rank-based Mahalanobis distance in Sect. 9.3. The function `smahal(·,·)` in the `DOS2` package creates the 131×2689 distance matrix. The first five rows and columns of `dmat` are given in Table 14.3.

Table 14.3 First five rows and columns of the 131×2689 distance matrix using the rank-based Mahalanobis distance

ℓ	Control 1	Control 2	Control 3	Control 4	Control 5
Treated 7	3.86	2.61	5.06	6.86	6.72
Treated 20	3.47	3.03	7.58	6.23	5.82
Treated 108	9.60	19.47	20.02	24.62	13.03
Treated 126	6.81	8.05	12.93	10.74	9.88
Treated 145	8.70	15.09	17.74	18.86	12.37

Notice that the 131 rows and 2689 columns are labeled with distinct numbers $\ell = 1, 2, \ldots, 2820 = 131 + 2689$

[4]Technically, the fitted probabilities in logit regression are invariant under affine transformations of the predictors.

Table 14.4 First five rows and columns of the 131 × 2689 distance matrix after adding the propensity score calipers

ℓ	Control 1	Control 2	Control 3	Control 4	Control 5
Treated 7	18.60	20.64	42.04	79.91	46.10
Treated 20	46.32	3.03	72.66	51.18	73.30
Treated 108	82.94	47.40	115.60	39.07	111.01
Treated 126	57.81	13.64	86.16	47.54	85.51
Treated 145	8.70	54.51	33.32	113.31	30.34

The caliper was at $0.2 \times \mathtt{sd}(\widehat{e}(\mathbf{x}))$. Only two of these 25 entries were unaffected by the caliper

```
> dmat<-smahal(zb,Xb)
> dim(dmat)
  [1] 131 2689
```

The second step adds the caliper on the propensity score in Sect. 9.4. The function `addcaliper(·,·,·)` in the `DOS2` package takes `dmat` and adds a penalty when the propensity scores differ by more than 0.2 times the standard deviation of $\widehat{e}(\mathbf{x})$. In the welder data in Chap. 9, the caliper was set at 0.5 times the standard deviation of $\widehat{e}(\mathbf{x})$, but aside from very small problems, a value of 0.1–0.25 is more common. The result is a new 131 × 2689 distance matrix. The first five rows and columns of the revised `dmat` are given in Table 14.4. Among these 25 entries, only two respected the caliper and did not incur a penalty.

```
> dmat<-addcaliper(dmat,zb,p)
> dim(dmat)
  [1] 131 2689
```

14.6 Constructing the Match

From the distance matrix, `dmat`, the match is obtained using Hansen's [17] `fullmatch` function in his `optmatch` package. After `optmatch` is installed, you must load it, either using the `packages menu` or with the command

```
> library(optmatch)
```

We wish to match ten controls to each senior with a deceased father. The `fullmatch` function needs to know the distance matrix, here `dmat`, the minimum number of controls, here 10, the maximum number of controls, here 10, and the fraction of controls to omit, here $(2689 - 10 \times 131)/2689 = 0.51283$. That is, matching 10-to-1 means 10×131 controls will be used, omitting $2689 - 10 \times 131$, which is 51%.

```
> 2689-10*131
  [1] 1379
> 1379/2689
  [1] 0.51283
```

With these quantities determined, we do the match:

```
> m<-fullmatch(dmat,min.controls=10,max.controls=10,
  omit.fraction=1379/2689)
```

After a few seconds, R places the match in m. There is an entry in m for each of the 2820 seniors.

```
> length(m)
  [1] 2820
```

The first ten entries in m are

```
> m[1:10]
   m.34 m.10 m.01 m.87 m.02 m.26 m.1 m.03 m.17 m.117
```

This says that the first senior of the 2820 seniors is in matched set #34 and the second senior is in matched set #10. The third senior was one of the 1379 unmatched controls; this is the meaning of the zero in m.01. The fourth senior is in matched set #87, the fifth is unmatched, and so on. The function matched(·) indicates who is matched. The first ten entries of matched(m) are

```
> matched(m)[1:10]
TRUE TRUE FALSE TRUE FALSE TRUE TRUE FALSE TRUE TRUE
```

which says, again, that the first two seniors were matched but the third was not, and so on. There are 1441 matched seniors, where $1441 = 131 \times 11$, because the 131 seniors with deceased fathers were each matched to ten controls, making 131 matched sets each of size 11.

```
> sum(matched(m))
  [1] 1441
> 1441/11
  [1] 131
```

The first three matched sets are in Table 14.5. The first match consists of 11 female high school seniors, neither black nor Hispanic, whose mothers had a high school education, with family incomes between $30,000 and $40,000, mostly with test scores between 59% and 77%. In the second matched set, incomes were lower but test scores were higher. And so on.

Three other matches were constructed: a 10-to-1 match for the period 1982–1983, after the Social Security Student Benefit Program had ended, a 1-to-1 match of seniors with fathers who were not deceased in 1982–1983 and in 1979–1981, and a variable match (Sect. 9.5) for all seniors whose fathers were deceased, with 54 matched sets containing one senior from 1982–1983 and between one and four seniors from 1979–1981; see Table 14.2. In producing the variable match, fullmatch parameters were set as follows: min.controls=1 to require at least one control, max.controls=4, to permit at most four controls, and omit.fraction=0 to require use of all 131 potential controls. The small variable match is included for illustration.

Table 14.5 The first three of 131 matched sets, each set containing one treated subject and ten matched controls

id	set	zb	faminc	incmiss	black	hisp	afqtpct	edmissm	edm	female
10	1380	1	3.22	0	0	0	61.92	0	2	1
350	1380	0	3.56	0	0	0	59.58	0	2	1
365	1380	0	3.60	0	0	0	61.24	0	2	1
465	1380	0	3.56	0	0	0	56.37	0	2	1
518	1380	0	3.79	0	0	0	67.01	0	2	1
550	1380	0	3.79	0	0	0	63.50	0	2	1
1294	1380	0	3.56	0	0	0	67.31	0	2	1
2072	1380	0	3.79	0	0	0	64.98	0	2	1
2082	1380	0	3.79	0	0	0	63.62	0	2	1
2183	1380	0	3.97	0	0	0	76.52	0	2	1
3965	1380	0	3.76	0	0	0	72.58	0	2	1
396	1381	1	2.37	0	0	0	88.51	0	2	1
2	1381	0	2.46	0	0	0	95.16	0	2	1
147	1381	0	2.27	0	0	0	77.09	0	2	1
537	1381	0	2.60	0	0	0	95.96	0	2	1
933	1381	0	2.85	0	0	0	96.11	0	2	1
974	1381	0	1.90	0	0	0	99.60	0	3	1
987	1381	0	2.13	0	0	0	81.18	0	2	1
1947	1381	0	2.05	0	0	0	91.45	0	3	1
2124	1381	0	2.30	0	0	0	72.40	0	2	1
2618	1381	0	2.21	0	0	0	68.92	0	2	1
3975	1381	0	2.37	0	0	0	90.74	0	2	1
3051	1382	1	3.41	0	0	1	62.87	0	1	0
606	1382	0	4.18	0	0	1	81.74	0	1	0
664	1382	0	4.39	0	0	1	91.57	0	1	0
884	1382	0	2.85	0	0	1	48.77	0	1	0
995	1382	0	3.13	0	0	1	55.12	0	1	0
1008	1382	0	3.32	0	0	1	51.61	0	1	0
1399	1382	0	3.44	0	0	1	90.02	0	2	0
2908	1382	0	3.80	0	0	1	80.28	0	1	0
3262	1382	0	3.79	0	0	1	57.28	0	1	0
3400	1382	0	2.93	0	0	1	80.88	0	2	0
3624	1382	0	3.32	0	0	1	74.08	0	1	1

Having constructed the match, the remaining work is done with standard R functions.

14.7 Checking Covariate Balance

Table 14.6 and Fig. 14.2 display the imbalance in the eight covariates and the propensity score in these four comparisons, before and after matching. For the variable match, the means in the control group are weighted; see Sect. 9.5.

The four pairs of two groups were quite different before matching but were much closer after matching. The family income of seniors with deceased fathers was lower, they were more often black, and their mothers had less education. Between 1979–1981 and 1982–1983, AFQT test scores decline. The imbalances are much smaller after matching.

14.8 College Outcomes

A treatment effect should enhance college attendance in the one group that received an incentive to attend, namely seniors in 1979–1981 whose fathers were deceased. If the other groups differ substantially from each other after matching for observed covariates, then this cannot be an effect of the treatment, and must indicate that matching has failed to render the groups comparable [10, 31, 35] and [43, Chapter 5]. Table 14.7 compares three measures of college attendance in four matched comparisons. In Table 14.7, the Mantel–Haenszel test and odds ratio are reported [16, §13.3] and percentages in the variable match are weighted; see Sect. 9.5.

In 1979–1981, when the Social Security Student Benefit Program provided tuition aid to students of a deceased Social Security beneficiary, seniors with deceased fathers were more likely to attend college and complete 1 year of college than were matched controls, with an odds ratio of about 1.65, but there is no sign of this in 1982–1983 after the program ended. Among seniors whose fathers were not deceased, college outcomes were slightly but not significantly better in 1982–1983 than among matched controls in 1979–1981. Seniors with deceased fathers in 1982–1983 ($n = 54$) and their variably matched seniors with deceased fathers in 1979–1981 ($n = 131$) did not differ significantly on any outcome, but the sample size is small, and the point estimates are not incompatible with reduced college attendance in 1982–1983, when the tuition benefit was withdrawn. There was no indication of a difference in completion of 4 years of college for any comparison.

Table 14.6 Balance on covariates before and after matching for four matched comparisons

Comparison	set	$\widehat{e}(\mathbf{x})$	faminc	incmiss	black	hisp	afqtpct	edmissm	edm	female
1979–1981	$\overline{x}_{tk}$	0.07	2.78	0.15	0.35	0.15	49.58	0.08	1.62	0.49
FD vs. FND	$\overline{x}_{cmk}$	0.06	2.77	0.15	0.34	0.15	49.10	0.04	1.61	0.50
	$\overline{x}_{ck}$	0.05	4.58	0.19	0.29	0.15	52.39	0.04	1.91	0.50
131 sets	sd_{bk}	0.67	0.71	0.11	0.13	0.02	0.10	0.19	0.33	0.03
10-to-1	sd_{mk}	0.09	0.00	0.00	0.02	0.03	0.02	0.19	0.02	0.02
1982–1983	$\overline{x}_{tk}$	0.08	2.31	0.30	0.46	0.15	37.41	0.04	1.56	0.46
FD vs. FND	$\overline{x}_{mck}$	0.07	2.53	0.33	0.41	0.13	39.13	0.03	1.62	0.49
	$\overline{x}_{ck}$	0.05	4.31	0.24	0.30	0.16	43.81	0.06	1.83	0.47
54 sets	sd_{bk}	0.76	0.80	0.12	0.35	0.04	0.22	0.11	0.34	0.02
10-to-1	sd_{mk}	0.18	0.09	0.08	0.11	0.06	0.06	0.03	0.07	0.05
FND	$\overline{x}_{tk}$	0.30	4.31	0.24	0.30	0.16	43.81	0.06	1.83	0.47
1982–1983	$\overline{x}_{cmk}$	0.30	4.33	0.24	0.29	0.15	44.06	0.05	1.81	0.47
vs 1979–1981	$\overline{x}_{ck}$	0.27	4.58	0.19	0.29	0.15	52.39	0.04	1.91	0.50
1038	sd_{bk}	0.37	0.09	0.12	0.01	0.03	0.30	0.10	0.09	0.06
pairs	sd_{mk}	0.02	0.01	0.00	0.01	0.02	0.01	0.05	0.03	0.00
FD	$\overline{x}_{tk}$	0.35	2.31	0.30	0.46	0.15	37.41	0.04	1.56	0.46
1982–1983	$\overline{x}_{cmk}$	0.33	2.49	0.25	0.43	0.15	38.55	0.06	1.65	0.43
vs 1979–1981	$\overline{x}_{ck}$	0.27	2.78	0.15	0.35	0.15	49.58	0.08	1.62	0.49
54 sets	sd_{bk}	0.67	0.24	0.35	0.23	0.01	0.43	0.20	0.09	0.05
variable	sd_{mk}	0.16	0.09	0.12	0.07	0.00	0.04	0.12	0.12	0.06

FD = father deceased and FND = father not deceased. For covariate k, $\overline{x}_{tk}$ is the mean in the first group in the comparison, $\overline{x}_{ck}$ is the mean in the second group in the comparison, and $\overline{x}_{mck}$ is the mean in the matched comparison group formed from the second group. As in Chap. 10, sd_{bk} is the standardized difference in means before matching and sd_{mk} is the standardized difference in means after matching; they have the same denominator but different numerators. The largest of 36 standardized differences before matching was 0.80, compared with 0.19 after matching

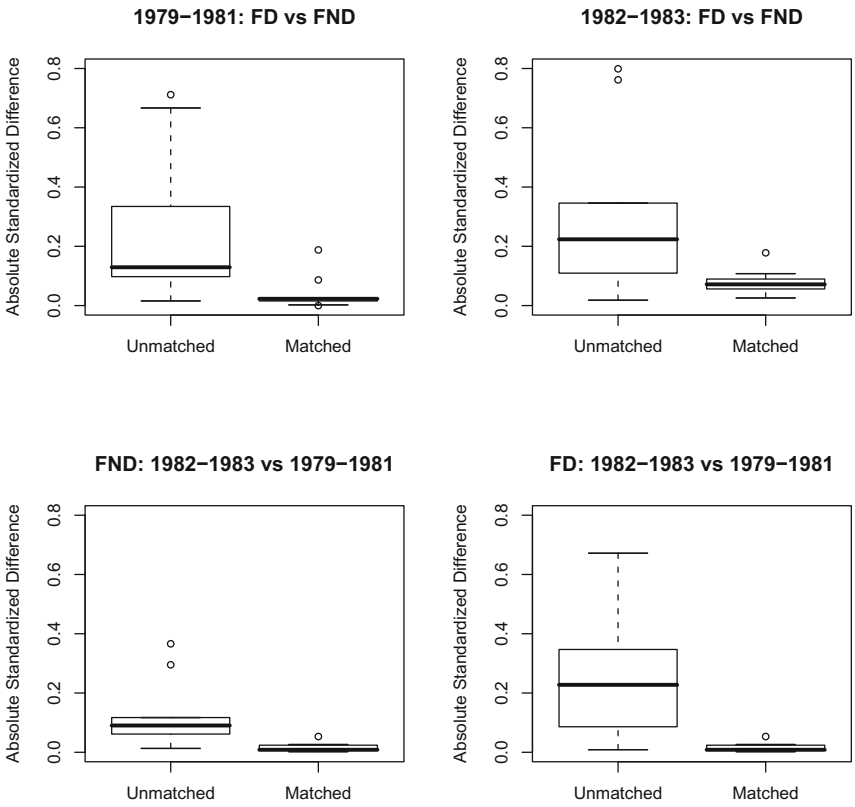

Fig. 14.2 Standardized differences in covariate means before and after matching in three matched comparisons. Each boxplot displays standardized differences in means, for the nine covariates and the propensity score

14.9 Further Reading

Susan Dynarski's [15] fine study is very much worth reading. The reanalysis presented here was intended just to illustrate many aspects of matching in R, not as a recommended analysis for a study of this form. Because of the several sample sizes in Dynarski's [15] study, several different forms of matching could be used with one data set. For a biomedical example, see [38].

Optimal matching is best done with dedicated software, for instance, the several R packages mentioned in Sect. 14.10. Some of the authors of these packages have written helpful articles that describe and illustrate use of the packages [17, 23, 27, 30]. These packages implement techniques discussed in earlier chapters but not illustrated in the current chapter, such as matching with fine balance or matching to strengthen an instrumental variable.

The mathematics of optimal matching in particular, and of combinatorial optimization in general, is interesting but extensive. A concise and elementary introduction to the optimal matching problem, or equivalently the optimal assignment problem, is given by Bertsekas [6]. His auction algorithm provides an especially tangible and familiar view of the competition among treated individuals for the same closest control. The auction algorithm literally holds an auction, selling controls to treated individuals, with prices that adjust to reflect competition for the same control. The technical concepts of duality and complementary slackness acquire familiar economic interpretations in the auction algorithm. For detailed discussion, see [8].

Two illustrations of the presentation of optimal matching in the medical literature are Kelz et al. [20] with 18,200 matched pairs, and Silber et al. [44] with 25,076 matched pairs.

14.10 Software

As noted in the abstract to this chapter, matching involves bookkeeping and efficient use of computer memory that are best handled by dedicated software for matching. This chapter has shown intermediate steps that software might automate. Dedicated software also implements other aspects of multivariate matching.

Ben Hansen's `optmatch` package in `R` performs various types of optimal matching, and it is the best choice for full matching [17, 18]. An `R` package by Bo Lu et al. [23], `nbpMatching`, performs nonbipartite matching or matching without groups. Cinar Kilcioglu and José Zubizarreta's `designmatch` package

Table 14.7 College outcomes by age 23: attended, completed a year, and completed 4 years

Outcome	Comparison	Group 1 %	Group 2 %	MH–odds ratio	p-value
Attend College	1979–1981: FD vs. FND	53	43	1.67	0.019
	1982–1983: FD vs. FND	35	37	0.90	0.853
	FND: 1982–1983 vs. 1979–1981	43	39	1.21	0.070
	FD: 1982–1983 vs. 1979–1981	35	46	0.63	0.264
Complete Any College	1979–1981: FD vs. FND	50	40	1.65	0.022
	1982–1983: FD vs. FND	31	34	0.87	0.798
	FND: 1982–1983 vs. 1979–1981	40	37	1.20	0.100
	FD: 1982–1983 vs. 1979–1981	31	36	0.81	0.369
Complete 4 Years	1979–1981: FD vs. FND	15	14	1.03	0.915
	1982–1983: FD vs. FND	9	12	0.70	0.630
	FND: 1982–1983 vs. 1979–1981	14	13	1.13	0.460
	FD: 1982–1983 vs. 1979–1981	9	12	0.88	0.805

In matched groups, the table gives percentages, the Mantel–Haenszel odds ratio, and the two-sided P-value from the Mantel–Haenszel test. Percents for the variable match are weighted

offers a variety of new tools for optimal matching [21, 48]. Because it uses mixed integer programming rather than network optimization, `designmatch` lacks a guarantee on the speed of computations, but it often performs competitively. Samuel Pimentel's `rcbalance` and `rcsubset` packages find an optimal match with fine or refined balance [27, 28]. Ruoqi Yu's `bigmatch` package solves very large optimal matching problems [47]. Ruoqi Yu's `DiPs` package implements matching with directional penalties [46]. Rigdon, Baiocchi, and Basu's [29] `R` package `nearfar` strengthens an instrumental variable using nonbipartite matching.

The `match` function in Ruoqi Yu's `DiPs` package is fairly basic, and it resembles the structure in this chapter; so, it is natural choice for a first attempt to use optimal matching. For instance, to try optimal matching with fine balance, run the example in the help-file for the `match` function in the `DiPs` package. The `DiPs` package contains other newer features [46], but the `match` function may be used without engaging these newer features. Do note that you must install and load the `optmatch` package to use the `DiPs` package.

14.11 Data

The `dynarski` data in the `R` package `DOS2` contains Dynarski's [15] data as used in this chapter, and the examples in its help-file reproduce some of the computations in this chapter. Additional data sets illustrating optimal matching are contained in several of the `R` packages discussed in Sect. 14.10.

Bibliography

1. Aitkin, M., Francis, B., Hinde, J., Darnell, R.: Statistical Modelling in R. Oxford University Press, New York (2009)
2. Angrist, J.D., Lavy, V.: Using Maimonides' rule to estimate the effect of class size on scholastic achievement. Q. J. Econ. **114**, 533–575 (1999)
3. Baiocchi, M., Small, D.S., Lorch, S., Rosenbaum, P.R.: Building a stronger instrument in an observational study of perinatal care for premature infants. J. Am. Stat. Assoc. **105**, 1285–1296 (2010)
4. Baiocchi, M., Small, D.S., Yang, L., Polsky, D., Groeneveld, P.W.: Near/far matching: a study design approach to instrumental variables. Health Serv. Outcomes Res. Method **12**, 237–253 (2012)
5. Bertsekas, D.P.: A new algorithm for the assignment problem. Math. Program **21**, 152–171 (1981)
6. Bertsekas, D.P.: The auction algorithm for assignment and other network flow problems: a tutorial. Interfaces **20**, 133–149 (1990)
7. Bertsekas, D.P.: Linear Network Optimization. MIT Press, Cambridge (1991)
8. Bertsekas, D.P.: Network Optimization: Continuous and Discrete Models. Athena Scientific, Belmont (1998)
9. Bertsekas, D.P., Tseng, P.: The relax codes for linear minimum cost network flow problems. An. Oper. Res. **13**, 125–190 (1988)

10. Campbell, D.T.: Factors relevant to the validity of experiments in social settings. Psychol. Bull. **54**, 297–312 (1957)
11. Card, D., Krueger, A.: Minimum wages and employment: a case study of the fast-food industry in New Jersey and Pennsylvania. Am. Econ. Rev. **84**, 772–793 (1994). http://www.irs.princeton.edu/
12. Chambers, J.: Software for Data Analysis: Programming with R. Springer, New York (2008)
13. Dalgaard, P.: Introductory Statistics with R. Springer, New York (2002)
14. Derigs, U. : Solving nonbipartite matching problems by shortest path techniques. Ann. Operat. Res. **13**, 225–261 (1988)
15. Dynarski, S.M.: Does aid matter? Measuring the effect of student aid on college attendance and completion. Am. Econ. Rev. **93**, 279–288 (2003)
16. Fleiss, J.L., Levin, B., Paik, M.C.: Statistical Methods for Rates and Proportions. Wiley, New York (2001)
17. Hansen, B.B.: Optmatch: flexible, optimal matching for observational studies. R News **7**, 18–24 (2007)
18. Hansen, B.B., Klopfer, S.O.: Optimal full matching and related designs via network flows. J. Comp. Graph. Stat. **15**, 609–627 (2006)
19. Ho, D., Imai, K., King, G., Stuart, E.A.: Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. Polit. Anal. **15**, 199–236 (2007)
20. Kelz, R.R., Sellers, M.M., Niknam, B.A., Sharpe, J.E., Rosenbaum, P.R., Hill, A.S., Zhou, H., Hochman, L.L., Bilimoria, K.Y., Itani, K., Romano, P.S., Silber, J.H.: A National comparison of operative outcomes of new and experienced surgeons. Ann. Surgery (2020). https://doi.org/10.1097/SLA.0000000000003388
21. Kilcioglu, C., Zubizarreta, J.R.: Maximizing the information content of a balanced matched sample in a study of the economic performance of green buildings. Ann. Appl. Stat. **10**, 1997–2020 (2016)
22. LaLonde, R.J.: Evaluating the econometric evaluations of training programs with experimental data. Am. Econ. Rev. **76**, 604–620 (1986)
23. Lu, B., Greevy, R., Xu, X., Beck, C.: Optimal nonbipartite matching and its statistical applications. Am. Stat. **65**, 21–30 (2011)
24. Maindonald, J., Braun, J.: Data Analysis and Graphics Using R. Cambridge University Press, New York (2001)
25. McCullagh, P. , Nelder, J.A. : Generalized Linear Models. Chapman and Hall/CRC, New York (1989)
26. Ming, K., Rosenbaum, P.R.: A note on optimal matching with variable controls using the assignment algorithm. J. Comput. Graph. Stat. **10**, 455–463 (2001)
27. Pimentel, S.D.: Large, sparse optimal matching with R package rcbalance. Obs. Stud. **2**, 4–23 (2016)
28. Pimentel, S.D., Kelz, R.R., Silber, J.H., Rosenbaum, P.R.: Large, sparse optimal matching with refined covariate balance in an observational study of the health outcomes produced by new surgeons. J. Am. Stat. Assoc. **110**, 515–527 (2015)
29. R Development Core Team.: R: a Language and Environment for Statistical Computing. R Foundation, Vienna (2019). http://www.R-project.org
30. Rigdon, J., Baiocchi, M., Basu, S.: Near-far matching in R: the nearfar package. J. Stat. Soft. **86**, 5. https://doi.org/10.18637/jss.v086.c05
31. Rosenbaum, P.R.: From association to causation in observational studies. J. Am. Stat. Assoc. **79**, 41–48 (1984)
32. Rosenbaum, P.R.: Permutation tests for matched pairs with adjustments for covariates. Appl. Stat. **37**, 401–411 (1988) (Correction: [36, §3])
33. Rosenbaum, P.R.: Optimal matching in observational studies. J. Am. Stat. Assoc. **84**, 1024–32 (1989)
34. Rosenbaum, P.R.: A characterization of optimal designs for observational studies. J. R. Stat. Soc. B **53**, 597–610 (1991)
35. Rosenbaum, P.R.: Stability in the absence of treatment. J. Am. Stat. Assoc. **96**, 210–219 (2001)

36. Rosenbaum, P.R.: Observational Studies (2nd ed.). Springer, New York (2002)
37. Rosenbaum, P.R.: Covariance adjustment in randomized experiments and observational studies (with Discussion). Stat. Sci. **17**, 286–327 (2002)
38. Rosenbaum, P.R.: Modern algorithms for matching in observational studies. Annu. Rev. Stat. Appl. **7**, 143–176 (2020). https://doi.org/10.1146/annurev-statistics-031219-041058
39. Rosenbaum, P.R., Rubin, D.B.: The central role of the propensity score in observational studies for causal effects. Biometrika **70**, 41–55 (1983)
40. Rosenbaum, P.R., Rubin, D.B.: Reducing bias in observational studies using subclassification on the propensity score. J. Am. Stat. Assoc. **79**, 516–524 (1984)
41. Rosenbaum, P.R., Rubin, D.B.: Constructing a control group by multivariate matched sampling methods that incorporate the propensity score. Am. Stat. **39**, 33–38 (1985)
42. Rubin, D.B.: Using multivariate matched sampling and regression adjustment to control bias in observational studies. J. Am. Stat. Assoc. **74**, 318–328 (1979)
43. Shadish, W.R., Cook, T.D., Campbell, D.T.: Experimental and Quasi-Experimental Designs for Generalized Causal Inference. Houghton-Mifflin, Boston (2002)
44. Silber, J.H., Rosenbaum, P.R., McHugh, M.D., Ludwig, J.M., Smith, H.L., Niknam, B.A., Even-Shoshan, O., Fleisher, L.A., Kelz, R.R., Aiken, L.H.: Comparison of the value of nursing work environments in hospitals across different levels of patient risk. JAMA Surg. **151**, 527–536 (2016)
45. Wooldridge, J.M.: Econometric Analysis of Cross Section and Panel Data. MIT Press, Cambridge (2002)
46. Yu, R., Rosenbaum, P.R.: Directional penalties for optimal matching in observational studies. Biometrics **75**(4), 1380–1390 (2019). https://doi.org/10.1111/biom.13098
47. Yu, R., Silber, J.H., Rosenbaum, P.R.: Matching methods for observational studies derived from large administrative databases. Stat. Sci. (2019)
48. Zubizarreta, J.R.: Using mixed integer programming for matching in an observational study of kidney failure after surgery. J. Am. Stat. Assoc. **107**, 1360–1371 (2012)

Part III
Design Sensitivity

Chapter 15
The Power of a Sensitivity Analysis and Its Limit

Abstract In an experiment, power and sample size calculations anticipate the outcome of a statistical test that will be performed when the experimental data are available for analysis. In parallel, in an observational study, the power of a sensitivity analysis anticipates the outcome of a sensitivity analysis that will be performed when the observational data are available for analysis. In both cases, it is imagined that the data will be generated by a particular model or distribution, and the outcome of the test or sensitivity analysis is anticipated for data from that model. Calculations of this sort guide many of the decisions made in designing a randomized clinical trial, and similar calculations may usefully guide the design of an observational study. In experiments, the power in large samples is used to judge the relative efficiency of competing statistical procedures. In parallel, the power in large samples of a sensitivity analysis is used to judge the ability of design features, such as those in Chap. 5, to distinguish treatment effects from bias due to unmeasured covariates. As the sample size increases, the limit of the power of a sensitivity analysis is a step function with a single step down from power 1 to power 0, where the step occurs at a value $\widetilde{\Gamma}$ of Γ called the design sensitivity. The design sensitivity is a basic tool for comparing alternative designs for an observational study.

15.1 The Power of a Test in a Randomized Experiment

15.1.1 What Is the Power of a Test?

The power of a statistical test anticipates the judgment that the test will issue when the test is put to use. Conceptually, the power of a test is the probability that the test will recognize that a false null hypothesis is indeed false and reject it. If the test will reject the null hypothesis when the P-value is less than or equal to, say, the conventional 0.05 level, then the power of the test is the probability that the P-value will be less than or equal to 0.05 when the null hypothesis is indeed false.

If the test is a valid test of its null hypothesis, and if the null hypothesis were true, the probability that the P-value would be less than or equal to 0.05 is itself less than or equal to 0.05. This is the defining property of a valid test or P-value. The

© Springer Nature Switzerland AG 2020

P. R. Rosenbaum, *Design of Observational Studies*, Springer Series in Statistics,
https://doi.org/10.1007/978-3-030-46405-9_15

P-value is unlikely to be less than or equal to 0.05 if the null hypothesis is true—it happens in only one experiment in 20. Having defined a test so that rejection of a true null hypothesis is improbable, we now want to make rejection of a false null hypothesis highly probable; that is, we would like to have a powerful test.

The power of a test depends upon many things. First and foremost, it depends upon what is true. If the null hypothesis is false, something else is true instead. If the null hypothesis is false but barely so, the power is likely to be little better than chance, 0.05. If the null hypothesis is far from the truth, the power is likely to be much higher. The power depends also upon the sample size, the duration of follow-up, the specifics of the experimental design, and the procedures used in statistical analysis.

Power is a basic tool in designing a randomized experiment. How many patients are needed? Well, calculate the power with 100, 200, and 300 patients. Is it better to study the survival of 200 patients for 5 years or 300 patients for 3 years? Calculate the power in each situation. The new treatment is expensive and difficult to apply. To reduce cost, the experimenter is considering randomizing one-third of the patients to treatment and two-thirds to control. Will this design be substantially inferior to randomizing half the patients to treatment and half to control? Calculate the power for both designs. There are two ways to measure the response, one precise but expensive, the other imprecise but inexpensive. If the inexpensive device is used, the money saved will permit a larger sample size. For fixed total cost, which is better, more precision with fewer patients, or less precision with more patients? Calculate the power in both situations. With 20 schools, it would be convenient to randomly assign ten schools to treatment, the rest to control, and to use the same treatment for the hundreds of students in each school. Is that a terrible idea? Would it be vastly better to randomize the many hundreds of students as individuals? Calculate the power for the two designs.

Power is also a basic tool in evaluating the statistical methods that are used to analyze the results of a randomized experiment. Faced with the same data from the same statistical models, two different tests will typically have different power. The power of two different tests in a variety of circumstances is often the basis for choosing which test to use. The t-test has slightly better power than the Wilcoxon test for data from a Normal distribution, but substantially inferior power for distributions with longer tails, which is a basic reason that the Wilcoxon test is preferred; its power is robust.

The remainder of this section briefly reviews the idea of power and its computation for Wilcoxon's signed rank statistic when used in a randomized experiment. The concept of power is important throughout Part III. The details of the computation of power are relevant to the details of the discussion in Part III, but the concepts of Part III should be accessible if the concept of power is clear.

15.1.2 A Pep Talk About Statistical Power

Design sensitivity and the power of a sensitivity analysis are simple but useful extensions of two other concepts, namely sensitivity analysis and the power of a statistical test. The concept of a sensitivity analysis was introduced in Sects. 3.4 and 3.5, and this might be a good moment to take a second look at those sections. The power of a statistical test will be very familiar to some readers and less so to others. If power is very familiar, please skip the next few paragraphs, continuing with the subsection "Computing power in a randomized experiment: The two steps."

For people to whom statistical power is less familiar, let me mention a common source of discomfort and misunderstanding. In testing a null hypothesis, we tentatively suppose that the null hypothesis is true, simply to push through the logic of hypothesis testing. That is, we say, "suppose the null hypothesis were true." If it were true, what is the chance of data as inconsistent with the null hypothesis as the data we observed? That chance is the P-value, and if the P-value is small, then we begin to doubt the supposition that the null hypothesis is true. That is the logic of hypothesis testing. In computing the power, we suppose that the null hypothesis is not true after all, and we ask about the probability of a small P-value when the null hypothesis is false. This second question, the one about power, can seem perfectly reasonable at one moment and incoherent the next, not unlike certain unstable optical illusions. The reason is that a power calculation supposes two contradictory things—that the null hypothesis is true and that it is not—and that can seem reasonable if the suppositions occur at different levels and incoherent if they collapse into a single contradictory supposition.

Rather than talk about power, let us talk about fishing. Suppose that I go fishing in a lake, but suppose there are no fish in the lake. Will I catch any fish? One perfectly reasonable answer is "no": if there are no fish in the lake, I will not catch any fish. Another answer is: if I supposed there were no fish in the lake, I would not go fishing in that lake; perhaps I would fish in some other lake, and then yes, perhaps I would catch a fish. The first answer was willing to suppose a situation in which I supposed something false, namely it supposed that I supposed there were fish in the lake, hence went fishing in the lake, but my supposition was false. In the second answer, the two levels of supposing have collapsed into one, and the idea of going fishing in a lake with no fish seems incoherent. The fish-in-lake situation is clear enough, but add some technical concepts, a few Greek symbols, a little jargon, and suppositional collapse can seem quite puzzling. In a power calculation, I suppose the null hypothesis is false, but I also suppose that I am ignorant of the fact that it is false, so I test it, and in my ignorance, in the process of testing the null hypothesis, I suppose it to be true. To think about the possibility of error, I must make suppositions at two levels: I must suppose something is true, and I must suppose that I can be ignorant of that truth. If the concept of power ever feels incoherent, go back to fish in a lake. In our ignorance, we scientists spend a lot of time fishing in lakes with no fish, and you need to get the basic concepts down.

In an analogous way, when in Sect. 15.2 we compute the power of a sensitivity analysis, we suppose that there is a treatment effect and there is no bias from unmeasured covariates, but we are ignorant of these facts, so we perform a sensitivity analysis when an omniscient investigator would not.[1] Again, there is the danger of misunderstanding from suppositional collapse. The sensitivity analysis entertains the possibility of no treatment effect and a bias of magnitude Γ, for several values of Γ, but the power of the sensitivity analysis is computed under the supposition of a treatment effect and no bias. The power of a sensitivity analysis asks: If the treatment worked, and there was no bias from unobserved covariates, then how would the sensitivity analysis turn out? This is parallel to the power of a test of no effect in a randomized experiment that asks: If the treatment worked, then how would the test of no effect turn out?

Think about what you want a sensitivity analysis to do. In an observational study, if you are looking at a treatment effect without bias, then you will not know this from the data, but you want the sensitivity analysis to report that the ostensible effects of the treatment are highly insensitive to unobserved bias. Perhaps that will happen, perhaps not. The power of the sensitivity analysis is the probability that it will happen when, indeed, there is a treatment effect without bias. Again, it is important to avoid suppositional collapse. You are imagining that the world is a certain way, but you are ignorant of this, and you are performing a statistical analysis, in this case a sensitivity analysis, in a state of ignorance. You want to know: Under what circumstances will a sensitivity analysis do what you want it to do? Under what circumstances will the analysis sort things out correctly?

Formulas for power are needed to produce numerical results and to prove assertions. If you are okay with formulas, then that is fine, but if not, take the assertions and numerical results as true, focus on conceptual issues, and move on quickly to Chap. 16 rather than getting bogged down here with a formula. For instance, right now, look ahead to Fig. 15.1 and realize that this figure is intelligible without reference to the technical material that precedes and produces it. Figure 15.1 says that, in a randomized experiment, the power is greater if the treatment effect is larger ($\tau = 1/2$ rather than $\tau = 1/4$), and the power increases to 1 as the sample size I increases. Figure 15.1 captures the way we think about randomized experiments. Figures 15.2 and 15.3 are also intelligible without reference to the technical material that precedes and produces them; look at them now. Figures 15.2 and 15.3 capture the key difference between power in a randomized experiment and power in an observational study. Figures 15.2 and 15.3 say that the power of a sensitivity analysis increases to 1 as the sample size I increases for small values

[1]The omniscient investigator would not test hypotheses. Then again, the omniscient investigator would not run a laboratory or collect data. So, the omniscient investigator would be unable to publish, and would be denied tenure. The omniscient investigator would protest the denial of tenure on the grounds that the university routinely grants tenure to investigators who remain fallible despite their expensive laboratories. The omniscient investigator would end up in an asylum. Be content that you are not an omniscient investigator, and instead try to objectively and publicly control the probability of error.

of Γ and decreases to 0 for large values of Γ, the dividing point being a quantity, $\widetilde{\Gamma}$, called the design sensitivity (which is $\widetilde{\Gamma} = 3.171$ in Fig. 15.2). Here, Γ is an aspect of your analysis and $\widetilde{\Gamma}$ is an aspect of the world; they cannot coexist at the same suppositional level. If you understand Figs. 15.1, 15.2, and 15.3, then you understand most of the concepts of this chapter, even if you do not know how to produce these figures.

15.1.3 Computing Power in a Randomized Experiment: The Two Steps

The power of Wilcoxon's signed rank statistic, T, in a randomized experiment is computed in two steps. First, for a particular α, conventionally $\alpha = 0.05$, one finds the critical value, ζ_α, such that the chance that $T \geq \zeta_\alpha$ if the null hypothesis is true is α.[2] The null hypothesis would be rejected in a one-sided 0.05 level test if $T \geq \zeta_{0.05}$ because the one-sided P-value would be less than or equal to 0.05. Second, one computes the power as the probability that $T \geq \zeta_\alpha$ when the null hypothesis is false and something else is true instead. The power is then the probability of rejecting the null hypothesis with a P-value less than or equal to 0.05 when the null hypothesis is, in fact, false.

15.1.4 Step 1: Determining the Critical Value Assuming the Null Hypothesis Is True

Consider the first step, the behavior of Wilcoxon's signed rank statistic, T, in a paired randomized experiment with I pairs when the null hypothesis of no treatment effect is true. It is convenient to assume there are no ties.[3] Recall from Sect. 2.3.3 that if the null hypothesis of no effect is true, then the expectation and variance of

[2]This statement is correct in concept, but it omits a small technical detail. The distribution of Wilcoxon's statistic, T, is discrete, so its distribution attaches dollops of probability to the possible integer values of T, namely $0, 1, \ldots, I\,(I+1)\,/2$. In consequence, there may be no value $\zeta_{0.05}$ such that $\Pr\left(T \geq \zeta_{0.05} | \mathscr{F}, \mathscr{Z}\right) = 0.05$. In consequence, the best feasible value, $\zeta_{0.05}$, is used, that is, the smallest value such that $\Pr\left(T \geq \zeta_{0.05} | \mathscr{F}, \mathscr{Z}\right) \leq 0.05$ when the null hypothesis is true. With this value, $\zeta_{0.05}$, the probability of rejection of a true hypothesis is less than or equal to 0.05—that is, the test has level 0.05—but the probability of false rejection may be slightly below 0.05—that is, the test has size slightly less than 0.05. In moderately large samples, I, the dollops of probability at integer values of T are very small, so this technical detail may be ignored for conceptual purposes. When an appeal is made to the central limit theorem as $I \to \infty$ to justify an approximate Normal null distribution for T, the technical detail disappears.

[3]Power will be computed for various continuous distributions of the responses, such as a Normal distribution, and for continuous distributions ties never occur; that is, the probability of a tie is zero. As has been seen several times, ties are a minor inconvenience, easy to address in practice,

T are given by

$$E\left(T \mid \mathscr{F}, \mathscr{Z}\right) = \frac{I\left(I+1\right)}{4}, \tag{15.1}$$

$$\operatorname{var}\left(T \mid \mathscr{F}, \mathscr{Z}\right) = \frac{I\left(I+1\right)\left(2I+1\right)}{24}, \tag{15.2}$$

and, after subtracting the expectation and dividing by the standard deviation, the distribution of T is well approximated by the standard Normal distribution if the number of pairs I is moderately large, that is, the standardized deviate converges in distribution to the standard Normal distribution,

$$\frac{T - E\left(T \mid \mathscr{F}, \mathscr{Z}\right)}{\sqrt{\operatorname{var}\left(T \mid \mathscr{F}, \mathscr{Z}\right)}} \longrightarrow N\left(0, 1\right) \quad \text{as} \quad I \to \infty\,; \tag{15.3}$$

see Lehmann [5, §3.2]. Recall also that the standard Normal cumulative distribution function is denoted $\Phi\left(\cdot\right)$ and its inverse, the quantile function, is denoted $\Phi^{-1}\left(\cdot\right)$, so that, for instance, $\Phi\left(1.65\right) = 0.95$ and $\Phi^{-1}\left(0.95\right) = 1.65$.

With $I = 100$ pairs, compute $E\left(T \mid \mathscr{F}, \mathscr{Z}\right) = I\left(I+1\right)/4$ as $100\left(100+1\right)/4$=2525, $\operatorname{var}\left(T \mid \mathscr{F}, \mathscr{Z}\right) = I\left(I+1\right)\left(2I+1\right)/24$ as $100\left(100+1\right)\left(2 \times 100+1\right)/24 = 84587.5$. Using (15.3) to approximate the probability that $T \geq 3005$, we calculate $\left(3005 - 2525\right)/\sqrt{84587.5} = 1.65$, and $\Pr\left(T \geq 3005 \mid \mathscr{F}, \mathscr{Z}\right) \doteq 1 - \Phi\left(1.65\right) = 1 - 0.95 = 0.05$. In Sect. 2.3.3, computations along these lines yielded approximate one-sided P-values for Wilcoxon's statistic. More generally, under the null hypothesis H_0, for any fixed ζ,

$$\Pr\left(T \geq \zeta \mid \mathscr{F}, \mathscr{Z}\right) \doteq 1 - \Phi\left\{\frac{\zeta - E\left(T \mid \mathscr{F}, \mathscr{Z}\right)}{\sqrt{\operatorname{var}\left(T \mid \mathscr{F}, \mathscr{Z}\right)}}\right\}. \tag{15.4}$$

The Normal approximation works increasingly well as the sample size increases, $I \to \infty$; however, $I = 100$ is quite adequate for most practical uses of (15.4).

For power calculations, we do not want the realized P-value computed in Sect. 2.3.3 from (15.4), but rather the quantile, ζ_α, of the distribution of T that corresponds with a P-value less than or equal to α, conventionally $\alpha = 0.05$. Rearranging (15.4) yields $\Pr\left(T \geq \zeta_\alpha \mid \mathscr{F}, \mathscr{Z}\right) \doteq \alpha$ if

$$\zeta_\alpha \doteq E\left(T \mid \mathscr{F}, \mathscr{Z}\right) + \Phi^{-1}\left(1-\alpha\right)\sqrt{\operatorname{var}\left(T \mid \mathscr{F}, \mathscr{Z}\right)}. \tag{15.5}$$

For instance, with $I = 100$ pairs and $\alpha = 0.05$, $\Phi^{-1}\left(1 - 0.05\right) = 1.645$, so the critical value $\zeta_{0.05}$ is computed as approximately $2525 + 1.645\sqrt{84587.5} = 3005$. Of course, (15.4) and (15.5) agree that $\Pr\left(T \geq 3005 \mid \mathscr{F}, \mathscr{Z}\right) \doteq 0.05$ for $I = 100$

but they tend to clutter theoretical arguments. As is typically done, ties are assumed absent until a circumstance arises in which it actually becomes important to address the issue.

pairs, but (15.5) reverses the computation in (15.4) determining ζ_α from α. With $I = 100$ pairs in a randomized experiment, we would reject the null hypothesis of no treatment effect in a one-sided $\alpha = 0.05$ level test if T turned out to be at least $\zeta_{0.05} = 3005$.

Wilcoxon's signed rank statistic T has the convenient property that the quantiles ζ_α of its null distribution are determined by α and I without reference to $\mathscr{F}$. See the discussion of "'distribution free" statistics in Sect. 2.3.3. In particular, the approximation (15.5) is computed from α and I using the standard Normal distribution $\Phi(\cdot)$ and (15.1) and (15.2).

15.1.5 Step 2: Determining the Power Assuming the Null Hypothesis Is False

Consider now the second step, determining the power as the probability that $T \geq \zeta_\alpha$ when the null hypothesis is false. If the null hypothesis is false, then something else is true, and the power depends upon what is, in fact, true. One way to proceed would be to specify in detail what is true, that is to specify $\mathscr{F}$, in particular to specify the responses, $\left(r_{Tij}, r_{Cij}\right)$, each subject ij would exhibit under treatment and under control. This specification would permit the calculation of the conditional power given $\mathscr{F}$, that is, the conditional probability that $T \geq \zeta_\alpha$ given $\mathscr{F}, \mathscr{Z}$. Somewhat more precisely, the conditional power is $\Pr(T \geq \zeta_\alpha | \mathscr{F}, \mathscr{Z})$ for some specific $\mathscr{F}$ for which there is a nonzero treatment effect. Computing the conditional power for $I = 100$ pairs of two subjects would entail specifying $\left(r_{Tij}, r_{Cij}\right)$ for two hundred individuals, which sounds like a thankless task. Instead, we imagine that $\mathscr{F}$ was generated by a stochastic model, and compute the unconditional power—invariably just called "the power"—as the probability that $T \geq \zeta_\alpha$ for all random $\mathscr{F}$'s generated by this model. That is, the (unconditional) power for Wilcoxon's statistic is $\Pr(T \geq \zeta_\alpha | \mathscr{Z})$ which is the expectation of $\Pr(T \geq \zeta_\alpha | \mathscr{F}, \mathscr{Z})$ over the specified model generating $\mathscr{F}$'s.

In principle, with a sufficiently fast computer, the problem of computing the power is now "solved." That is, (i) simulate an $\mathscr{F}$ from the specified model, (ii) randomize treatments in the I pairs, (iii) compute Wilcoxon's T from the treated-minus-control differences in observed responses, Y_i, (iv) score a 1 if $T \geq \zeta_\alpha$, score a 0 if $T < \zeta_\alpha$, and (v) repeat steps (i)–(iv) many times, estimating the power as the proportion of 1's. In practice, it is often helpful to work a little harder and obtain an expression for the power, $\Pr(T \geq \zeta_\alpha | \mathscr{Z})$.

15.1.6 A Simple Case: Constant Effect with Random Errors

This definition of power is quite flexible, and its flexibility is put to use in later chapters, but to make things tangible, consider the simplest nontrivial case, though nonetheless an important case. In the simplest case, the treated-minus-control differences in observed responses, $Y_i = (Z_{i1} - Z_{i2})(R_{i1} - R_{i2})$, are independent and identically distributed (iid) observations drawn from a distribution $F(\cdot)$. For instance, a simple model to generate $\mathscr{F}$ has differences in responses to control, $r_{Ci1} - r_{Ci2}$, that are iid observations from a Normal distribution with expectation 0 and variance ω^2—that is, $r_{Ci1} - r_{Ci2} \sim_{iid} N\left(0, \omega^2\right)$—and an additive constant treatment effect, $r_{Tij} - r_{Cij} = \tau$ for all ij. Then, in a randomized experiment, the treated-minus-control difference in observed responses, $Y_i = (Z_{i1} - Z_{i2})(R_{i1} - R_{i2})$, is $Y_i = \tau + (Z_{i1} - Z_{i2})(r_{Ci1} - r_{Ci2})$, where given $\mathscr{F}$ and $\mathscr{Z}$, the quantity $Z_{i1} - Z_{i2}$ is ± 1 with equal probabilities, so $Y_i \sim_{iid} N\left(\tau, \omega^2\right)$. The power in this case is the chance that $T \geq \zeta_\alpha$ when Wilcoxon's signed rank statistic is computed from I independent observations drawn from $N\left(\tau, \omega^2\right)$.

When computing the power for differences Y_i that are sampled independently from a continuous distribution $F(\cdot)$, Lehmann [5, §4.2] gives an approximation to the power of Wilcoxon's signed rank statistic, T. The approximation appeals to the central limit theorem twice as the sample size I increases, $I \to \infty$, once to obtain the critical value ζ_α in step 1 as has already been done, and the second time to approximate the behavior of T when the null hypothesis is false. For $Y_i \sim_{iid} F(\cdot)$, define $p = \Pr(Y_i > 0)$, $p_1^{'} = \Pr(Y_i + Y_{i'} > 0)$ and $p_2^{'} = \Pr(Y_i + Y_{i'} > 0 \text{ and } Y_i + Y_{i''} > 0)$ with $i < i' < i''$. Then Lehmann shows that the expectation μ_F of Wilcoxon's T when the null hypothesis is false and the Y_i are sampled independently from $F(\cdot)$ is

$$\mu_F = \frac{I(I-1)\,p_1^{'}}{2} + Ip\,, \tag{15.6}$$

the variance of T is

$$\sigma_F^2 = I(I-1)(I-2)\left(p_2^{'} - p_1^{'\,2}\right) \tag{15.7}$$

$$+\frac{I(I-1)}{2}\left\{2\left(p - p_1^{'}\right)^2 + 3p_1^{'}\left(1 - p_1^{'}\right)\right\} + Ip(1-p)\,, \tag{15.8}$$

and that the central limit theorem yields approximate power as $\Pr(T \geq \zeta_\alpha | \mathscr{Z}) \doteq 1 - \Phi\{(\zeta_\alpha - \mu_F)/\sigma_F\}$.[4]

[4]This is expression (4.28) in Lehmann [5, §4.2] except for the exclusion of the continuity correction.

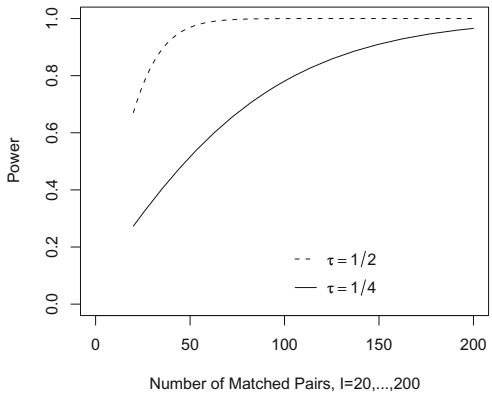

Fig. 15.1 Approximate power of Wilcoxon's signed rank test, T, in a paired randomized experiment when the treatment has an additive constant effect τ and the treated-minus-control differences in observed responses Y_i are independently sampled from a Normal distribution, $N(\tau, 1)$, with expectation τ and variance 1, for samples of size $I = 20, 21, \ldots, 200$ pairs

The required quantities p, $p_1^{'}$, and $p_2^{'}$ are determined by the distribution $F(\cdot)$. For the Normal distribution, p and $p_1^{'}$ have simple expressions, while $p_2^{'}$ may be obtained by numerical integration or simulation. For any $F(\cdot)$, it is easy to estimate p, $p_1^{'}$, and $p_2^{'}$ very precisely by simulating one very large sample of Y_i's drawn from $F(\cdot)$.[5]

For instance, if $Y_i \sim_{iid} N\left(\tau, \omega^2\right)$ with $\tau = \frac{1}{4}$ and $\omega = 1$ (or more generally with $\tau/\omega = \frac{1}{4}$), then $p = 0.599$, $p_1^{'} = 0.638$, $p_2^{'} = 0.482$. For $I = 100$, we found earlier that $\zeta_{0.05} \doteq 3005$. For $I = 100$ pairs, $\mu_F = 3217.16$, $\sigma_F^2 = 76400.5$, and $\Pr(T \geq \zeta_\alpha | \mathscr{Z}) \doteq 1 - \Phi\{(\zeta_\alpha - \mu_F)/\sigma_F\} = 0.78$. In other words, if the null hypothesis of no effect is false in a randomized experiment in such a way that the treated-minus-control differences in observed outcomes are Normal with expectation equal to one quarter of their standard deviation, then the probability that the null hypothesis will be rejected at the 0.05 level in a one-sided test is about 78% when there are 100 pairs. In the same way, with $\tau = \frac{1}{2}$ and $\omega = 1$ (or more generally with $\tau/\omega = \frac{1}{2}$) and $I = 100$ pairs, $\mu_F = 3834.67$, $\sigma_F^2 = 57041.2$, and $\Pr(T \geq \zeta_\alpha | \mathscr{Z}) \doteq 1 - \Phi\{(\zeta_\alpha - \mu_F)/\sigma_F\} = 0.999$.

Figure 15.1 plots the power of Wilcoxon's test for $Y_i \sim_{iid} N(\tau, 1)$ with $\tau = \frac{1}{4}$ or $\tau = \frac{1}{2}$ for $I = 20, \ldots 200$ pairs. With $I = 50$ pairs, the power is above 95% if

[5]The probability $p_1^{'} = \Pr(Y_i + Y_{i'} > 0)$ will turn out to be the most important of the three probabilities. Notice that in (15.6), the quantity $p_1^{'}$ dominates the expectation of T, because $p_1^{'}$ is multiplied by $I(I+1)/2$ while p is multiplied by I, so in large samples, as $I \to \infty$, the dominant role is played by $p_1^{'}$. Looking back to Sect. 2.5, we see that $p_1^{'}$ is the expectation of $H_{\mathscr{I}}$ for $\mathscr{I} = \{i, i'\}$. As seen in Sect. 2.5, the quantity $H_{\mathscr{I}}$ is highly interpretable, and the interpretation carries over to $p_1^{'}$.

the effect is larger, $\tau = \frac{1}{2}$, but the power is a coin toss, about 50%, if the effect is smaller, $\tau = \frac{1}{4}$.

15.2 Power of a Sensitivity Analysis in an Observational Study

15.2.1 What Is the Power of a Sensitivity Analysis?

As seen in Chap. 2, if Wilcoxon's signed rank test is used to test the null hypothesis of no treatment effect in a randomized experiment, it yields a single P-value because we know the distribution of treatment assignments that was created by the random assignment of treatments. As seen in Chap. 3, in an observational study, randomization is not used to assign treatments, so this single P-value lacks a justification or warrant; it would be valid under the naïve model, but naïveté is not justification for an inference. Recall from Sects. 3.4 and 3.5 that for any given magnitude $\Gamma \geq 1$ of departure from random assignment in (3.13), an interval $[P_{\min}, P_{\max}]$ of possible P-values may be determined; see Table 3.2. The study is judged sensitive to unmeasured biases of magnitude Γ if the corresponding $P_{\min}$ is small and $P_{\max}$ is large, for instance, $P_{\min} \leq 0.01$ and $P_{\max} > 0.05$. In Table 3.2, the study by Werfel et al. [17] was highly insensitive to unmeasured biases, specifically insensitive to biases of magnitude $\Gamma = 4$ but sensitive to $\Gamma = 5$.

If an observational study were actually free from unmeasured biases—if the naïve model (3.5)–(3.8) of Sect. 3.3 were true—then we would not know this from the observable data. We might see that the treated-minus-control differences in observed responses, Y_i, were typically positive, but we would be uncertain whether this reflected an effect caused by the treatment or some bias in the way treatments were assigned or a combination of the two. The best we could hope to say in this favorable situation is that the treatment appears to work and that this appearance is highly insensitive to unmeasured biases, that is, only a very large value of Γ could produce this appearance from bias alone. Speaking informally, the power of a sensitivity analysis is the probability that our hopes will be realized.

Somewhat more precisely, define the "favorable situation" as a treatment that actually causes meaningful effects in an observational study that is free of unmeasured biases. Specifically, in the favorable situation, the null hypothesis of no effect is false, and not merely false but the treatment effect is large enough to be interesting and substantively important. Also, in the favorable situation, the study is free of unmeasured biases, people who look comparable are comparable, in the sense that (3.5)–(3.8) are true. That is, the treated-minus-control differences in responses, Y_i, are not centered at zero because the treatment worked in the absence of bias. If we were in the favorable situation, we would not know it from the observable data. It is in this favorable situation that we would like to report a high degree of insensitivity to unmeasured biases.

In testing the null hypothesis of no treatment effect at level α, conventionally $\alpha = 0.05$, rejection of the null hypothesis is insensitive to a bias of magnitude Γ if $P_{\max} \leq \alpha$; see, again, the example in Table 3.2. In the favorable situation, the power of a sensitivity analysis conducted with a specified α and Γ is the probability that $P_{\max} \leq \alpha$. That is, if the treatment had an effect and there is no unmeasured bias, then the power of the sensitivity analysis for specified α and Γ is the probability that we will be able to say that a bias of magnitude Γ would not lead to acceptance of the null hypothesis of no effect when tested at level α. In Table 3.2, it was found that $P_{\max} \leq 0.05$ for $\Gamma = 4$; the power is the probability of such an event computed before the data are examined. For $\Gamma = 1$, this is the power of a test of no effect in a randomized experiment as computed in Sect. 15.1.

15.2.2 Computing the Power of a Sensitivity Analysis: The Two Steps

As in Sect. 15.1, the computations leading to the power of a sensitivity analysis are done in two steps. As in Sect. 15.1, the first step is to compute a critical value, now $\zeta_{\Gamma,\alpha}$, so that $\Pr\left(T \geq \zeta_{\Gamma,\alpha} \middle| \mathscr{F}, \mathscr{Z}\right) \leq \alpha$ for all treatment assignment probabilities that satisfy (3.16)–(3.18) when the null hypothesis of no treatment effect is true.[6] If we observed $T \geq \zeta_{\Gamma,\alpha}$, then we would report that the upper bound on the one-sided P-value for testing no effect is at most α for all biases of magnitude at most Γ; see again Table 3.2. The second step is to compute the probability that $T \geq \zeta_{\Gamma,\alpha}$ in the "favorable situation" in which the treatment has an effect and there is actually no bias from unmeasured covariates. It is in this favorable situation that we hope to report a high degree of insensitivity to unmeasured bias, and if $T \geq \zeta_{\Gamma,\alpha}$, then we will be able to report that rejection of the null hypothesis is insensitive to a bias of magnitude Γ.

As in Sect. 15.1, the computation of the critical value is little more than a rearrangement of the computation of the P-value for a one-sided test. Specifically, rearranging the computation (3.23) in Sect. 3.5 yields

$$\zeta_{\Gamma,\alpha} \doteq E\left(\overline{\overline{T}} \middle| \mathscr{F}, \mathscr{Z}\right) + \Phi^{-1}(1-\alpha)\sqrt{\operatorname{var}\left(\overline{\overline{T}} \middle| \mathscr{F}, \mathscr{Z}\right)} \tag{15.9}$$

so that using (3.19) and (3.20) yields

$$\zeta_{\Gamma,\alpha} \doteq \frac{\Gamma}{1+\Gamma} \cdot \frac{I(I+1)}{2} + \Phi^{-1}(1-\alpha)\sqrt{\frac{\Gamma}{(1+\Gamma)^2}\frac{I(I+1)(2I+1)}{6}}. \tag{15.10}$$

[6]To be precise, we seek the smallest $\zeta_{\Gamma,\alpha}$ such that $\Pr\left(T \geq \zeta_{\Gamma,\alpha} \middle| \mathscr{F}, \mathscr{Z}\right) \leq \alpha$.

For $\Gamma = 1$ in a randomized experiment, expressions (15.9) and (15.10) reduce to (15.5).

Continuing the numerical illustration from Sect. 15.1, with $I = 100$ pairs, $\alpha = 0.05$, and $\Gamma = 2$, the critical value $\zeta_{2,\,0.05}$ is approximately

$$\frac{2}{1+2} \cdot \frac{100\,(100+1)}{2} + 1.645\sqrt{\frac{2}{(1+2)^2}\frac{100\,(100+1)\,(2\cdot 100+1)}{6}} \tag{15.11}$$

or $\zeta_{2,\,0.05} = 3817.7$, which is noticeably higher than $\zeta_{0.05} = \zeta_{1,\,0.05} \doteq 3005$ for a randomized experiment $\Gamma = 1$ in Sect. 15.1. In words, with $I = 100$ pairs, if $T \geq 3817.7$, then the maximum possible P-value is ≤ 0.05 for all treatment assignment probabilities that satisfy (3.16)–(3.18) with $\Gamma = 2$. This completes the first step.

15.2.3 Step 2: Determining the Power When the Null Hypothesis Is False and There Is No Unobserved Bias

Again, we wish to compute the power of the sensitivity analysis—the chance that $T \geq \zeta_{\Gamma,\alpha}$—in the favorable situation in which there is no bias from unmeasured covariates and the treatment has an effect. It is in this situation that we hope to report that the treatment appears to work and the appearance is insensitive to moderately large biases.

More precisely, the naïve model (3.5)–(3.8) is assumed to hold, or equivalently (3.16)–(3.18) holds with $\Gamma = 1$, and the data $\mathscr{F}$ were generated by a stochastic model with a treatment effect, and we wish to determine $\Pr\left(T \geq \zeta_{\Gamma,\alpha} \middle| \mathscr{Z}\right)$ averaging $\Pr\left(T \geq \zeta_{\Gamma,\alpha} \middle| \mathscr{F}, \mathscr{Z}\right)$ over the model that generates $\mathscr{F}$. This is, of course, identical to the second step in Sect. 15.1 for a randomized experiment except the critical value, $\zeta_{\Gamma,\alpha}$, has changed. In consequence, the second step is performed in the same way as in Sect. 15.1 with the new critical value. As in Sect. 15.1, the second step may in general be performed by simulating many sets of data $\mathscr{F}$ from the model, determining the proportion with $T \geq \zeta_{\Gamma,\alpha}$, and in particular cases the power of the sensitivity analysis may be determined analytically.

Continuing the numerical example, if $Y_i \sim_{iid} N\left(\tau, \omega^2\right)$, $i = 1, \ldots, I = 100$ with $\tau = \frac{1}{2}$ and $\omega = 1$, then as in Sect. 15.1, $\mu_F = 3834.67$, $\sigma_F^2 = 57041.2$. To compute $\Pr\left(T \geq \zeta_{\Gamma,\alpha} \middle| \mathscr{Z}\right)$ for $\Gamma = 2$ and $\alpha = 0.05$, we find in the first step that $\zeta_{2,\,0.05} = 3817.7$. In the second step, we approximate $\Pr\left(T \geq \zeta_{\Gamma,\alpha} \middle| \mathscr{Z}\right)$ by $1 - \Phi\left\{\left(\zeta_{\Gamma,\alpha} - \mu_F\right)/\sigma_F\right\}$ which is $1 - \Phi\left\{(3817.7 - 3834.67)/\sqrt{57041.2}\right\} = 0.528$. If the treatment had an effect of $\tau = \frac{1}{2}$ and there is no bias from unobserved covariates, there is a 53% chance that a sensitivity analysis will yield an upper bound on the P-value that is at most 0.05 when performed with $\Gamma = 2$. This contrasts with 99.9% power in a randomized experiment, $\Gamma = 1$, as calculated in Sect. 15.1.

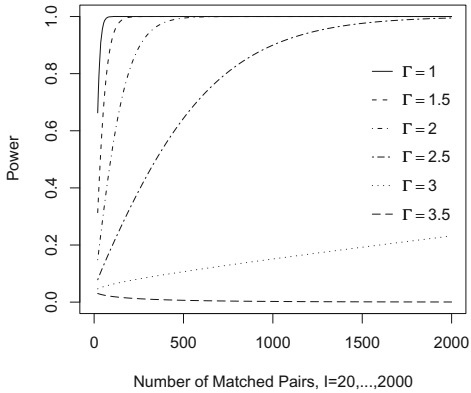

Fig. 15.2 Approximate power of a sensitivity analysis when there is no bias from unmeasured covariates, the treatment has an additive constant effect $\tau = 0.5$ and the treated-minus-control differences in observed responses Y_i are independently sampled from a Normal distribution, $N(\tau, 1)$, with expectation τ and variance 1, for samples of size $I = 20, 21, \ldots, 2000$ pairs. The several curves give the power for several values of Γ. The power tends to 1 as I increases for $\Gamma < 3.171$ and it tends to 0 for $\Gamma > 3.171$

15.2.4 A First Look At the Power of a Sensitivity Analysis

Figure 15.2 shows the power in the "favorable situation" with $Y_i \sim_{iid} N\left(\frac{1}{2}, 1\right)$, $i = 1, \ldots, I$ for $I = 20, 21, \ldots, 2000$ and for $\Gamma = 1, 1.5, 2, 2.5, 3, 3.5$. Again, the "favorable situation" is characterized by a treatment effect without bias, so it is the situation where we hope to report insensitivity to small and moderate unmeasured biases. The rather special case in Fig. 15.2 illustrates several general properties of the power of a sensitivity analysis. The power is higher in a randomized experiment, where Γ is known to equal 1, because there is no need to take account of bias from nonrandom treatment assignment. The power degrades as Γ increases. The most striking feature of Fig. 15.2 is that the power is increasing in the sample size I for small Γ but is decreasing in I for $\Gamma = 3.5$. There is, in fact, a value, $\widetilde{\Gamma}$, such that the power $\Pr\left(T \geq \zeta_{\Gamma,\alpha} \middle| \mathscr{Z}\right)$ tends to 1 as $I \to \infty$ for all $\Gamma < \widetilde{\Gamma}$ and $\Pr\left(T \geq \zeta_{\Gamma,\alpha} \middle| \mathscr{Z}\right)$ tends to 0 as $I \to \infty$ for all $\Gamma > \widetilde{\Gamma}$. In the situation in Fig. 15.2, the design sensitivity is $\widetilde{\Gamma} = 3.171$. Because $3 < \widetilde{\Gamma} = 3.171$, the power for $\Gamma = 3$ will eventually rise to 1 as $I \to \infty$ even though it has not yet reached 30% at $I = 2000$ in Fig. 15.2. In a sufficiently large sample, I, the "favorable situation" with $Y_i \sim_{iid} N\left(\frac{1}{2}, 1\right)$, $i = 1, \ldots, I$ can be distinguished from bias of magnitude $\Gamma < \widetilde{\Gamma} = 3.171$, but not from biases of magnitude $\Gamma > \widetilde{\Gamma} = 3.171$. Think of this statement as the quantitative replacement for the statement "association does not imply causation." In the "favorable situation" with $Y_i \sim_{iid} N\left(\frac{1}{2}, 1\right)$, $i = 1, \ldots, I$, the association between treatment Z and response Y can be distinguished from biases less than $\widetilde{\Gamma}$ but not from biases greater than $\widetilde{\Gamma}$.

It is time to consider the design sensitivity, $\widetilde{\Gamma}$, in greater detail.

15.3 Design Sensitivity

15.3.1 A First Look at Design Sensitivity

In many simple settings, the power of a sensitivity analysis tends to 1 as the sample size I increases if the analysis is performed with a sufficiently small value of Γ, but the power tends to 0 if the analysis is performed with a sufficiently large value of Γ. The switch occurs at a value, $\widetilde{\Gamma}$, called the design sensitivity [7]. In the situation in Fig. 15.2, the design sensitivity was $\widetilde{\Gamma} = 3.171$. It would be natural to seek a design that results in a larger design sensitivity and to avoid designs that result in a smaller design sensitivity. It would be natural to appraise the performance of the methods in Chap. 5 in terms of their impact on design sensitivity.

Figure 15.2 plotted the power as a function of the sample size for several values of Γ. Although such a plot resembles the familiar plot of power for a randomized experiment (e.g., Fig. 15.1), it is actually less instructive than plotting the power as a function of Γ for several sample sizes I. Figure 15.3 plots power against Γ for samples of size $I = 200, 2000, 20{,}000$ in the "favorable situation" in which there is a treatment effect but no unmeasured bias, with $Y_i \sim_{iid} N(\tau, 1)$, $i = 1, \ldots, I$, for $\tau = \frac{1}{2}$ and $\tau = 1$.

The first thing to notice about Fig. 15.3 is that, as the sample size I increases, the power viewed as a function of Γ is tending to a step function with a single step down from power 1 to power 0, with the step located at the design sensitivity, $\widetilde{\Gamma}$. For $\tau = \frac{1}{2}$, the step is at $\widetilde{\Gamma} = 3.171$, whereas for $\tau = 1$ the step is at $\widetilde{\Gamma} = 11.715$. For all sample sizes, I, there is negligible power for $\Gamma > \widetilde{\Gamma}$.

The six power curves in Fig. 15.3 refer to three sample sizes, $I = 200, 2000, 20{,}000$, and two design sensitivities, $\widetilde{\Gamma} = 3.171$ (for $\tau = \frac{1}{2}$) and $\widetilde{\Gamma} = 11.715$ (for $\tau = 1$). In this specific 3×2 array of power curves, the difference in design sensitivities matters more than the difference in sample sizes, although sample size does have a substantial effect on power. For instance, at $\Gamma = 5$, the power is close to 1 for $I = 200$ pairs with $\tau = 1$ but the power is close to zero with $I = 20{,}000$ pairs and $\tau = \frac{1}{2}$. A bias of magnitude $\Gamma = 5$ could explain the observed association between treatment and outcome if $\tau = \frac{1}{2}$ no matter how large the sample size becomes, whereas a mere $I = 200$ pairs could eliminate bias of magnitude $\Gamma = 5$ as a plausible explanation if $\tau = 1$. A smaller study with a larger effect ($I = 200$, $\tau = 1$) is very likely to be less sensitive to unmeasured bias than a vastly larger study with a smaller effect ($I = 20000$, $\tau = \frac{1}{2}$). It is often asserted that, in observational studies, biases from unmeasured covariates constitute a greater source of uncertainty than does a limited sample size, and Fig. 15.3 is one quantitative expression of this assertion.

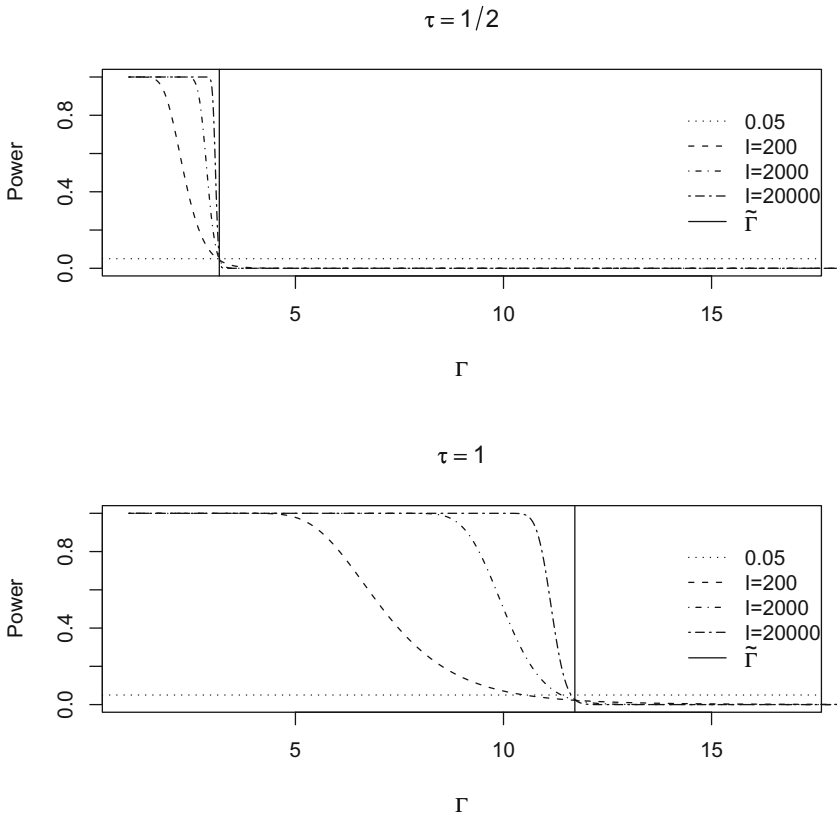

Fig. 15.3 Approximate power of a sensitivity analysis when there is no bias from unmeasured covariates, the treatment has an additive constant effect τ and the treated-minus-control differences in observed responses Y_i are independently sampled from a Normal distribution, $N(\tau, 1)$, with expectation τ and variance 1, for samples of size $I = 200$, 2000, and 20000 pairs. The solid vertical line is at the design sensitivity, $\tilde{\Gamma}$. The power tends to 1 as I increases for $\Gamma < \tilde{\Gamma}$ and tends to 0 for $\Gamma > \tilde{\Gamma}$. The power is computed for a 0.05 level test, and the horizontal dotted line is at 0.05

15.3.2 A Formula for Design Sensitivity

Design sensitivity is a general concept applicable to many statistics, not just Wilcoxon's signed rank statistic, and to many sampling situations, not just matched pairs [7]. However, in the case of Wilcoxon's signed rank statistic for matched pairs, there is a simple explicit formula for the design sensitivity [3, 15], which is given in Proposition 15.1. Notice that this explicit form uses the quantity $p_1^{'} = \Pr(Y_i + Y_{i'} > 0)$, which was encountered in Sect. 15.1 in connection with

Lehmann's formula for the approximate power of Wilcoxon's signed rank statistic when applied in a randomized experiment.[7]

Proposition 15.1 *In I pairs matched exactly for observed covariates, $\mathbf{x}_{ij}$, suppose that the treated-minus-control matched pair differences in outcomes, Y_i, $i = 1, \ldots, I$, are independently sampled from a distribution $F(\cdot)$ and there is no bias from unmeasured covariates in the sense that treatments are actually assigned by the naïve model (3.5)–(3.8). For a sensitivity analysis applied to Wilcoxon's signed rank statistic T in a one-tailed α-level test of no effect with $\alpha > 0$, the power of the sensitivity analysis $\Pr\left(T \geq \zeta_{\Gamma,\alpha} \middle| \mathscr{Z}\right)$ satisfies*

$$\Pr\left(T \geq \zeta_{\Gamma,\alpha} \middle| \mathscr{Z}\right) \to 1 \text{ as } I \to \infty \text{ for } \Gamma < \widetilde{\Gamma} \tag{15.12}$$

and

$$\Pr\left(T \geq \zeta_{\Gamma,\alpha} \middle| \mathscr{Z}\right) \to 0 \text{ as } I \to \infty \text{ for } \Gamma > \widetilde{\Gamma}, \tag{15.13}$$

where

$$\widetilde{\Gamma} = \frac{p_1^{'}}{1 - p_1^{'}} \text{ with } p_1^{'} = \Pr\left(Y_i + Y_{i'} > 0\right), \quad i \neq i'. \tag{15.14}$$

The proof is given in the appendix to this chapter.

15.3.3 Computing Design Sensitivity with Additive Effects and iid Errors

If the Y_i, $i = 1, \ldots, I$, are independently sampled from a distribution $F(\cdot)$, then it is always easy to determine $p_1^{'} = \Pr\left(Y_i + Y_{i'} > 0\right)$, $i \neq i'$, and hence also $\widetilde{\Gamma}$, by simulation. In the case of an additive effect τ with Normal or Cauchy errors, explicit formulas are available. Recall that $\Phi(\cdot)$ is the standard Normal cumulative distribution, $N(0, 1)$, and write $\Upsilon(\cdot)$ for the cumulative Cauchy distribution. As in [5, pages 166–167], if $Y_i \sim_{iid} N\left(\tau, \omega^2\right)$, then $(Y_i - \tau)/\omega \sim_{iid} N(0, 1)$ and $(Y_i + Y_{i'} - 2\tau)/\sqrt{2\omega^2} \sim N(0, 1)$, so

$$\Pr\left(Y_i + Y_{i'} > 0\right) = \Pr\left(\frac{Y_i + Y_{i'} - 2\tau}{\omega\sqrt{2}} > \frac{-2\tau}{\omega\sqrt{2}}\right) = \Phi\left(\sqrt{2}\frac{\tau}{\omega}\right). \tag{15.15}$$

[7] See also Note 5 in Sect. 15.1.

For instance, if $Y_i \sim_{iid} N\left(\frac{1}{2}, 1\right)$, as in Figs. 15.2 and 15.3, then $\Pr(Y_i + Y_{i'} > 0) = \Phi\left(\sqrt{2} \cdot \frac{1}{2}\right) = 0.76025$ and $\widetilde{\Gamma} = 0.76025/(1 - 0.76025) = 3.171$. If $(Y_i - \tau)/\omega \sim \Upsilon(\cdot)$, then $\{(Y_i - \tau)/\omega + (Y_{i'} - \tau)/\omega\}/2 \sim \Upsilon(\cdot)$, so

$$\Pr(Y_i + Y_{i'} > 0) = \Pr\left(\frac{Y_i + Y_{i'}}{2} > 0\right) = \Pr\left(\frac{Y_i + Y_{i'} - 2\tau}{2\omega} > \frac{-2\tau}{2\omega}\right) = \Upsilon\left(\frac{\tau}{\omega}\right). \tag{15.16}$$

For instance, if $Y_i - \frac{1}{2} \sim \Upsilon(\cdot)$, then $\Pr(Y_i + Y_{i'} > 0) = \Upsilon\left(\frac{1}{2}\right) = 0.64758$ and $\widetilde{\Gamma} = 0.64758/(1 - 0.64758) = 1.838$.

15.4 Summary

As seen in Fig. 15.3 and Proposition 15.1, no matter how large the sample size I becomes, a study will be sensitive to bias of magnitude Γ if $\Gamma > \widetilde{\Gamma}$ where $\widetilde{\Gamma}$ is the design sensitivity. More precisely, when testing the null hypothesis of no effect, the power of a sensitivity analysis will tend to 1 as $I \to \infty$ for all $\Gamma < \widetilde{\Gamma}$ and to 0 for all $\Gamma > \widetilde{\Gamma}$. In this sense, $\widetilde{\Gamma}$ measures the limit of the ability of a particular study design to distinguish a treatment effect from a bias due to failure to control for an unmeasured covariate u_{ij}. For this reason, the design sensitivity $\widetilde{\Gamma}$ is a natural numerical measure for comparing design strategies that hope to distinguish treatment effects from unmeasured biases, such as the strategies discussed in Chap. 5. The remainder of Part III will make such comparisons.

15.5 Further Reading

The remainder of Part III is further reading for this chapter. Design sensitivity was introduced in [7] and is developed further for study design in [2, 3, 8, 9, 11, 12, 14–16].

Design sensitivity is closely connected to Bahadur efficiency, the concept of large-sample efficiency introduced by Raghu Raj Bahadur; e.g., [1]. The most familiar concept of large-sample efficiency is Pitman efficiency, which considers a sequence of ever smaller treatment effects as the sample size increases, $I \to \infty$. The difficulty with a straightforward interpretation of Pitman efficiency in sensitivity analyses is that small treatment effects are invariably sensitive to small biases. If we focus on small—hence, relatively unimportant—treatment effects, we are led away from situations in which there is some hope of finding results that are insensitive to small biases. In contrast, Bahadur efficiency considers how quickly, as the sample size increases, as $I \to \infty$, we can confidently distinguish a nontrivial

treatment effect and zero treatment effect. The Bahadur efficiency is compatible with the design sensitivity: specifically, the Bahadur efficiency of a sensitivity analysis gradually declines to zero as Γ increases towards the design sensitivity $\widetilde{\Gamma}$. Moreover, the Bahadur efficiency provides useful information about the behavior of sensitivity analyses for $\Gamma < \widetilde{\Gamma}$; that is, information that $\widetilde{\Gamma}$ cannot provide; see Sect. 19.5 and [13].

Appendix: Technical Remarks and Proof of Proposition 15.1

This appendix contains a proof of Proposition 15.1 in Sect. 15.3. The formula and proof [3, 15] for Wilcoxon's signed rank statistic are a special case of a general result [7, §3] with the attractive feature that the design sensitivity $\widetilde{\Gamma}$ has an explicit closed form (15.14).

The approximate power is $\Pr\left(T \geq \zeta_{\Gamma,\alpha} \middle| \mathscr{Z}\right) \doteq 1 - \Phi\left\{\left(\zeta_{\Gamma,\alpha} - \mu_F\right)/\sigma_F\right\}$, where μ_F is the expectation of T in the "favorable situation." An intuitive, heuristic derivation [7, 15] uses the notion that if $\zeta_{\Gamma,\alpha} > \mu_F$, then in large samples, as $I \to \infty$, the signed rank statistic T eventually will be less than $\zeta_{\Gamma,\alpha}$, whereas if $\zeta_{\Gamma,\alpha} < \mu_F$, then T will eventually be greater than $\zeta_{\Gamma,\alpha}$. So the heuristic equates $\zeta_{\Gamma,\alpha}$ in (15.10) and μ_F in (15.6) and solves for Γ. It is slightly better to equate $\zeta_{\Gamma,\alpha}/\mu_F$ to 1, let $I \to \infty$, ignore terms that are small compared with I^2, and solve for Γ,

$$\frac{\zeta_{\Gamma,\alpha}}{\mu_F} \doteq \frac{\frac{\Gamma I(I+1)}{2(1+\Gamma)} + \Phi^{-1}(1-\alpha)\sqrt{\frac{\Gamma I(I+1)(2I+1)}{6(1+\Gamma)^2}}}{I(I-1)\,p_1^{'}/2 + Ip} \tag{15.17}$$

$$\doteq \frac{\Gamma}{p_1^{'}(1+\Gamma)}, \tag{15.18}$$

where $1 = \Gamma/\left\{p_1^{'}(1+\Gamma)\right\}$ has solution $\widetilde{\Gamma} = p_1^{'}/\left(1 - p_1^{'}\right)$ as in (15.14). The heuristic is attractive because it uses only the expectation μ_F of T in the favorable situation, so the calculation is simple. Alas, the heuristic is "heuristic" precisely because the power depends on $\left(\zeta_{\Gamma,\alpha} - \mu_F\right)/\sigma_F$, not on $\zeta_{\Gamma,\alpha} - \mu_F$, and the heuristic ignores σ_F, hoping that things will work out. Indeed, things do work out, as the proof that follows demonstrates [3]. The general result in [7, §3] shows that the heuristic calculation works quite generally, but the proof that follows is self-contained and does not depend upon the general result.

Proof For large I, the power, $\Pr\left(T \geq \zeta_{\Gamma,\alpha} \middle| \mathscr{Z}\right)$, is approximately equal to $1 - \Phi\left\{\left(\zeta_{\Gamma,\alpha} - \mu_F\right)/\sigma_F\right\}$, so the power tends to 1 as $I \to \infty$ if $\left(\zeta_{\Gamma,\alpha} - \mu_F\right)/\sigma_F \to -\infty$ and the power tends to 0 as $I \to \infty$ if $\left(\zeta_{\Gamma,\alpha} - \mu_F\right)/\sigma_F \to \infty$. Write $\kappa = \Gamma/(1+\Gamma)$. Using the expressions for μ_F and σ_F^2 from (15.6) and (15.7)

in Sect. 15.1 and the expression for $\zeta_{\Gamma,\alpha}$ from (15.10) in Sect. 15.2 yields

$$\frac{\zeta_{\Gamma,\alpha} - \mu_F}{\sigma_F} = \tag{15.19}$$

$$\frac{\kappa I\left(I+1\right)/2 + \Phi^{-1}\left(1-\alpha\right)\sqrt{\kappa\left(1-\kappa\right)I\left(I+1\right)\left(2I+1\right)/6} - I\left(I-1\right)p_1^{'}/2 - Ip}{\sqrt{I\left(I-1\right)\left(I-2\right)\left(p_2^{'} - p_1^{'\,2}\right) + \frac{I(I-1)}{2}\left\{2\left(p-p_1^{'}\right)^2 + 3p_1^{'}\left(1-p_1^{'}\right)\right\} + Ip\left(1-p\right)}}. \tag{15.20}$$

Expression (15.20) looks untidy at first. Notice, however, that every term in the numerator and denominator of (15.20) involves I, and $I \to \infty$. So divide the numerator and denominator of (15.20) by $I\sqrt{I}$, and let $I \to \infty$. The last term in the numerator, $-Ip$, becomes $-Ip/\left(I\sqrt{I}\right) \to 0$ as $I \to \infty$, whereas the first term in the numerator, $\kappa I\left(I+1\right)/2$, becomes $\kappa I\left(I+1\right)/\left(2I\sqrt{I}\right) \approx \sqrt{I}\kappa/2$ in the sense that their ratio is tending to 1 as $I \to \infty$. Continuing in this way, we find that

$$\frac{\zeta_{\Gamma,\alpha} - \mu_F}{\sigma_F} \approx \frac{\sqrt{I}\left(\kappa - p_1^{'}\right)}{2\sqrt{p_2^{'} - p_1^{'\,2}}} \quad \text{as} \quad I \to \infty, \tag{15.21}$$

in the sense that the ratio of the left and right sides of (15.21) tends to 1 as $I \to \infty$ provided $\kappa \neq p_1^{'}$. If $\kappa > p_1^{'}$, then (15.21) tends to ∞ as $I \to \infty$, so the power tends to 0. If $\kappa < p_1^{'}$, then (15.21) tends to $-\infty$ as $I \to \infty$, so the power tends to one. Moreover, $\kappa = \Gamma/\left(1+\Gamma\right) > p_1^{'}$ if and only if $\Gamma > p_1^{'}/\left(1-p_1^{'}\right)$.

Bibliography

1. Bahadur, R.R.: Rates of convergence of estimates and test statistics. Ann. Math. Stat. **38**, 303–24 (1967)
2. Ertefaie, A., Small, D.S., Rosenbaum, P.R.: Quantitative evaluation of the trade-off of strengthened instruments and sample size in observational studies. J. Am. Stat. Assoc. **113**, 1122–1134 (2018)
3. Heller, R., Rosenbaum, P.R., Small, D.: Split samples and design sensitivity in observational studies. J. Am. Stat. Assoc. **104**, 1090–1101 (2009)
4. Hodges, J.L., Lehmann, E.L.: Estimates of location based on ranks. Ann. Math. Stat. **34**, 598–611 (1963)
5. Lehmann, E.L.: Nonparametrics. Holden Day, San Francisco (1975). Reprinted: Springer, New York (2006)
6. Rosenbaum, P.R.: Hodges-Lehmann point estimates of treatment effect in observational studies. J. Am. Stat. Assoc. **88**, 1250–1253 (1993)
7. Rosenbaum, P.R.: Design sensitivity in observational studies. Biometrika **91**, 153–164 (2004)

8. Rosenbaum, P.R.: Heterogeneity and causality: unit heterogeneity and design sensitivity in observational studies. Am. Stat. **59**, 147–152 (2005)
9. Rosenbaum, P.R.: What aspects of the design of an observational study affect its sensitivity to bias from covariates that were not observed? Festshrift for Paul W. Holland. ETS, Princeton (2009)
10. Rosenbaum, P.R.: A new U-statistic with superior design sensitivity in matched observational studies. Biometrics **67**, 1017–1027 (2011)
11. Rosenbaum, P.R.: Nonreactive and purely reactive doses in observational studies. In: Berzuini, C., Dawid, A.P., Bernardinelli, L. (eds.) Causality: Statistical Perspectives and Applications, pp. 273–289. Wiley, New York (2012)
12. Rosenbaum, P.R.: Impact of multiple matched controls on design sensitivity in observational studies. Biometrics **69**, 118–127 (2013)
13. Rosenbaum, P.R.: Bahadur efficiency of sensitivity analyses in observational studies. J. Am. Stat. Assoc. **110**, 205–217 (2015)
14. Rosenbaum, P.R.: Using Scheffé projections for multiple outcomes in an observational study of smoking and periodontal disease. Ann. App. Stat. **10**, 1147–1471 (2016)
15. Small, D., Rosenbaum, P.R.: War and wages: the strength of instrumental variables and their sensitivity to unobserved biases. J. Am. Stat. Assoc. **103**, 924–933 (2008)
16. Small, D.S., Cheng, J., Halloran, M.E., Rosenbaum, P.R.: Case definition and design sensitivity. J. Am. Stat. Assoc. **108**, 1457–1468 (2013)
17. Werfel, U., Langen, V., Eickhoff, I., Schoonbrood, J., Vahrenholz, C., Brauksiepe, A., Popp, W., Norpoth, K.: Elevated DNA single-strand breakage frequencies in lymphocytes of welders. Carcinogenesis **19**, 413–418 (1998)

Chapter 16
Heterogeneity and Causality

Abstract Before R.A. Fisher introduced randomized experimentation, the literature on causal inference emphasized reduction of heterogeneity of experimental units. To what extent is heterogeneity relevant to causal claims in observational studies when random assignment of treatments is unethical or infeasible?

16.1 J.S. Mill and R.A. Fisher: Reducing Heterogeneity or Introducing Random Assignment

In his *System of Logic: Principles of Evidence and Methods of Scientific Investigation*, John Stuart Mill [13] proposed "four methods of experimental inquiry," including the "method of difference":

> If an instance in which the phenomenon … occurs and an instance in which it does not … have every circumstance save one in common … [then] the circumstance [in] which alone the two instances differ is the … cause or a necessary part of the cause … [Mill wanted] "two instances … exactly similar in all circumstances except the one" [under study] [13, III, §8]

Notice Mill's emphasis on a complete absence of heterogeneity: "have every circumstance save one in common;" that is, on treated and control units that are identical but for the treatment. In the modern biology laboratory, nearly identical, genetically engineered mice are compared under treatment and control; this is a modern expression of Mill's "method of difference." It is clear from the quote that Mill believed, rightly or wrongly, that heterogeneity of experimental units is directly relevant to causal claims, and does not refer simply to reducing the standard error of an estimate.

Ronald Fisher [3, Chapter 2] took a starkly different view. In 1935, in Chapter 2 of his *Design of Experiments*, Fisher introduced randomized experimentation for the first time in book form, discussing his famous experiment of the lady tasting tea. Fisher was directly critical of the "method of difference":

© Springer Nature Switzerland AG 2020

P. R. Rosenbaum, *Design of Observational Studies*, Springer Series in Statistics,
https://doi.org/10.1007/978-3-030-46405-9_16

> It is not sufficient remedy to insist that "all the cups must be exactly alike" in every respect except that to be tested. For this is a totally impossible requirement in our example, and equally in all other forms of experimentation … These are only examples of the differences probably present; it would be impossible to present an exhaustive list of such possible differences … because [they] … are always strictly innumerable. When any such cause is named, it is usually perceived that, by increased labor and expense, it could be largely eliminated. Too frequently it is assumed that such refinements constitute improvements to the experiment … [3, page 18]

In the first omission, "…," in this quote, Fisher discussed the many ways two cups of tea may differ.

Fisher is, of course, engaged in an enormously important task: he is introducing the logic of randomized experimentation to a broad audience for the first time. Moreover, it would be reasonable to say that Fisher was correct and Mill was wrong in certain critical respects. In a randomized clinical trial conducted in a hospital, the patients are heterogeneous and not much can be done about it. There is no opportunity to replace the hospital's patients by genetically engineered, nearly identical patients. And yet it is possible to randomly assign treatments to heterogeneous patients and draw valid causal inferences in just the manner that Fisher advocated.

Despite this, one senses that there is at least something that is sensible in Mill's method of difference, something that is sensible in the fanatical efforts of the basic science laboratory to eliminate heterogeneity, in the use of nearly identical mice. One senses that Fisher, in his understandable enthusiasm for his new method, has gone just a tad too far in his criticism of Mill's drive to eliminate heterogeneity. Indeed, the issue may be particularly relevant in observational studies where random assignment of treatments is either unethical or infeasible.

Some care is required in discussing heterogeneity. Heterogeneity is itself heterogeneous; there are several kinds of heterogeneity. In the biology laboratory, it is often wise to use several different strains or species of genetically engineered, nearly identical laboratory animals, making sure that each strain or species is equally represented in treated and control groups, to verify that any conclusion is not a peculiarity of a single strain. It is uncontrolled rather than controlled heterogeneity that is the target for reduction or elimination. Controlled heterogeneity has numerous uses.[1]

See Paul Holland's [8] essay "Statistics and causal inference" for further contrasts of the views of Mill and Fisher.

[1]Recall the related discussion of Bitterman's concept of "control by systematic variation" in Sect. 5.2.2, where some factor is demonstrated to be irrelevant by systematically varying that factor.

16.2 A Larger, More Heterogeneous Study Versus a Smaller, Less Heterogeneous Study

16.2.1 Large I or Small σ: Which is Better?

To explore the issue raised in Sect. 16.1, consider the following simple situation. There are I matched pairs in an observational study with treated-minus-control differences in outcomes Y_i, $i = 1, \ldots, I$. Because it is an observational study, not a randomized experiment, we cannot assume that matching for observed covariates has removed all bias from nonrandom assignment—we cannot assume the naïve model of Chap. 3, and will report a sensitivity analysis. Although we cannot know this from the observed data, the situation is, in fact, the "favorable situation," in which there is a treatment effect and the matching has succeeded in removing bias, so the naïve model is correct and treatment assignment is effectively randomized within matched pairs; see Sect. 15.2. In this favorable situation, the investigator hopes to report that the treatment appears to be effective and that appearance is insensitive to small and moderate biases. Indeed, the situation is simpler still: the treatment has an additive, constant effect, $\tau = r_{Tij} - r_{Cij}$, so that $Y_i = \tau + (2Z_{i1} - 1)(r_{Ci1} - r_{Ci2})$; see Sect. 2.4.1. Moreover, the $r_{Ci1} - r_{Ci2}$ are independent and identically distributed observations drawn from a continuous distribution symmetric about zero; see Sect. 15.1. Because this is the favorable situation, $2Z_{i1} - 1 = \pm 1$, each with probability $\frac{1}{2}$ independently of $r_{Ci1} - r_{Ci2}$, so $Y_i - \tau$ itself has this same continuous distribution symmetric about zero.

The investigator faces a choice between a larger study with more heterogeneous responses or a smaller study with less heterogeneous responses, both in the "favorable situation." In Sect. 16.1, Mill would have advocated the smaller, less heterogeneous study. Is there any merit to Mill's claim? The heterogeneity here refers to heterogeneity that remains after matching for observed covariates, that is, heterogeneity within pairs; heterogeneity between pairs is not at issue. Specifically, the investigator faces the following admittedly stylized choice: observe either $4I$ pairs with additive effect τ and $(r_{Ci1} - r_{Ci2})/\omega \sim F(\cdot)$, where $F(\cdot)$ is a continuous distribution symmetric about zero, or alternatively observe I pairs with additive effect τ and $(r_{Ci1} - r_{Ci2})/(\omega/2) \sim F(\cdot)$. In words, the choice is between $4I$ pairs with dispersion ω or I pairs with dispersion $\omega/2$. The choice is stylized in the following sense. If $F(\cdot)$ were the standard Normal distribution, then the sample mean difference, $\overline{Y} = (1/I)\sum_{i=1}^{I} Y_i$, would be Normally distributed $\overline{Y} \sim N\left\{\tau, \omega^2/(4I)\right\}$ with expectation τ and variance $\omega^2/(4I)$ in both the larger, more heterogeneous study and the smaller, less heterogeneous study. If ω were known in a randomized experiment, the larger, more heterogeneous study and the smaller, less heterogeneous study would barely be worth distinguishing, because the sufficient statistic, $\overline{Y}$, has the same distribution in both studies.

Of course, this is not a randomized experiment. Does that matter for this choice? If so, how does it matter?

16.2.2 A Simulated Example

Figure 16.1 depicts a simulated example of the choice between a smaller, less heterogeneous study (SL) with $Y_i \sim N\left\{\tau,\ (\omega/2)^2\right\}$ for $i = 1, \ldots, I = 100$, and a larger, more heterogeneous study (LM) with $Y_i \sim N\left\{\tau,\ \omega^2\right\}$ for $i = 1, \ldots, I = 400$. In Fig. 16.1, $\tau = 1/2$, $\omega = 1$. The boxplots for SL and LM have 100 and 400 pairs, respectively.

If SL and LM were analyzed as if they were randomized experiments, the inferences would be very similar. In SL, the mean difference is $\overline{Y} = 0.487$ with estimated standard error 0.054, while in LM the mean difference is $\overline{Y} = 0.485$ with estimated standard error 0.049; however, the true standard error is 0.05 in both SL and LM. Using Wilcoxon's signed rank statistic to test the null hypothesis of no effect yields a very small P-value in both SL and LM, less than 10^{-10}. The Hodges–Lehmann point estimate $\widehat{\tau}$ of τ is $\widehat{\tau} = 0.485$ for SL and $\widehat{\tau} = 0.489$ for LM. The 95% confidence interval from the randomization distribution of Wilcoxon's statistic is $[0.374,\ 0.600]$ from SL and $[0.390,\ 0.587]$ from LM. If the choice were between two randomized experiments with the distributions SL and LM yielding Fig. 16.1, there would be little reason to prefer one over the other.

Suppose, however, that SL and LM came from observational studies, so the behavior of the Y_i might reflect either a treatment effect or an unmeasured bias or a combination of the two. How sensitive are the conclusions from SL and LM to departures from the naïve model (3.5)–(3.8) that underlies the inferences in the previous paragraph?

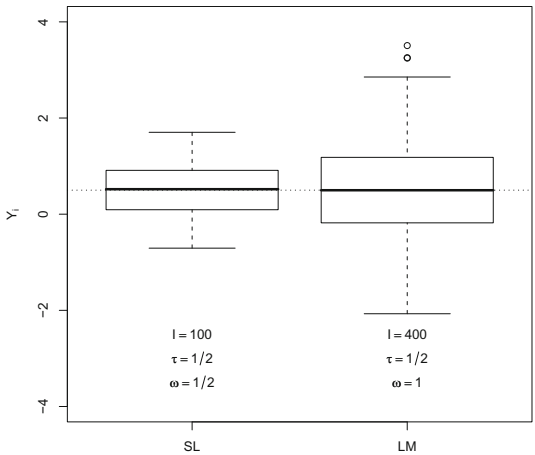

Fig. 16.1 A simulated example of the choice between a smaller, less heterogeneous study (SL) and a larger, more heterogeneous study (LM). In SL there are $I = 100$ independent matched pair differences, Y_i, that are Normal with expectation τ and standard deviation $\omega = 1/2$. In LM there are $I = 400$ independent matched pair differences, Y_i, that are Normal with expectation τ and standard deviation $\omega = 1$. The horizontal dotted line is at $1/2$

Table 16.1 Sensitivity analysis for the larger, more heterogeneous study (LM) and the smaller, less heterogeneous study (SL)

Γ	1	2	3	4	5
LM	$<10^{-10}$	0.00046	0.39	0.97	1.00
SL	$<10^{-10}$	0.000016	0.0021	0.022	0.083

Upper bounds on the one-sided P-value from Wilcoxon's signed rank test when testing the null hypothesis of no treatment effect are given. Although the randomization inferences are similar ($\Gamma = 1$), the smaller, less heterogeneous study is much less sensitive to bias from unmeasured covariates ($\Gamma \geq 3$)

Table 16.2 Sensitivity analysis for the larger, more heterogeneous study (LM) and the smaller, less heterogeneous study (SL)

	$\Gamma = 1$ $\widehat{\tau}$	$\Gamma = 2$ $[\widehat{\tau}_{\text{min}}, \widehat{\tau}_{\text{max}}]$	$\Gamma = 2$ $\widehat{\tau}_{\text{max}} - \widehat{\tau}_{\text{min}}$
LM	0.489	[0.19, 0.79]	0.60
SL	0.485	[0.32, 0.66]	0.34

For $\Gamma = 1$, the table gives the value of the Hodges–Lehmann point estimate $\widehat{\tau}$ of the treatment effect, τ. For $\Gamma = 2$, the table gives the interval of possible point estimates, $[\widehat{\tau}_{\text{min}}, \widehat{\tau}_{\text{max}}]$, and the length of that interval, $\widehat{\tau}_{\text{max}} - \widehat{\tau}_{\text{min}}$. For a bias of magnitude $\Gamma = 2$, the range of possible point estimates is much longer for the larger, more heterogeneous study than for the smaller, less heterogeneous study

Table 16.1 displays two sensitivity analyses, one for LM, the other for SL, giving the upper bound on the one-sided P-value for testing the hypothesis of no treatment effect using Wilcoxon's signed rank statistic; see Sect. 3.4. As noted above, the randomization inferences ($\Gamma = 1$) are quite similar for LM and SL. In sharp contrast, the smaller, less heterogeneous study, SL, is much less sensitive to bias from an unmeasured covariate. A bias of $\Gamma = 3$ could produce a boxplot similar to the LM boxplot in Fig. 16.1 (the upper bound on the P-value is 0.39), but a bias of $\Gamma = 3$ is very unlikely to produce the boxplot for SL (the upper bound on the P-value is 0.0021). To put this in context, SL is just slightly more sensitive to unmeasured bias than Hammond's [5] study of heavy smoking as a cause of lung cancer (see [16, §4]), one of the least sensitive observational studies, whereas LM is much more sensitive.

In parallel with Table 3.4, Table 16.2 displays two sensitivity analyses, one for LM, the other for SL, giving the interval of possible point estimates, $[\widehat{\tau}_{\text{min}}, \widehat{\tau}_{\text{max}}]$, of the treatment effect τ. For $\Gamma = 1$, the interval is a point, namely the Hodges–Lehmann point estimate, $\widehat{\tau}_{\text{min}} = \widehat{\tau}_{\text{max}} = \widehat{\tau}$, and it is about the same for LM and SL. For $\Gamma = 2$, the interval for LM, namely [0.19, 0.79], is considerably longer than the interval for SL, namely [0.32, 0.66].

In Tables 16.1 and 16.2, the smaller, less heterogeneous study is better than the larger, more heterogeneous study, better in the sense of being less sensitive to unmeasured biases. It is important to keep in mind that Fig. 16.1 depicts two

"favorable situations," that is, treatment effects without unmeasured biases. Because these are observational studies, the investigator does not know when she is in the "favorable situation," and so she cannot assert that Fig. 16.1 depicts effects, not biases. The investigator can, however, assert that a bias of magnitude $\Gamma = 4$ is too small to explain away as unmeasured bias the ostensible effect in the smaller, less heterogeneous study, but could not assert this about the larger, more heterogeneous study.

16.2.3 Power Comparisons with Normal, Logistic, and Cauchy Errors

Table 16.3 contrasts the larger, more heterogeneous study (LM) and the smaller, less heterogeneous study (SL) in terms of the power of the sensitivity analysis. Table 16.3 refers to a one-sided, 0.05 level test of the null hypothesis of no treatment effect. The power is the probability that the upper bound on the one-sided P-value is at most 0.05. The power is calculated as in Sect. 15.2.

The power of the randomization test ($\Gamma = 1$) is similar for LM and SL, but for larger Γ, particularly for $\Gamma = 2$, the power is higher for the smaller, less heterogeneous study (SL). The pattern seen in Table 16.1 is not a peculiarity of one simulation; rather, it is the anticipated pattern based on a comparison of the powers of LM and SL.

16.2.4 Design Sensitivity

Proposition 15.1 and expressions (15.15) and (15.16) may be used to determine the design sensitivity $\widetilde{\Gamma}$ of the larger, more heterogeneous study (LM) and the smaller,

Table 16.3 Power of a sensitivity analysis for a larger, more heterogeneous study (LM, $4I$ pairs, $\omega = 1$), and a smaller, less heterogeneous study, (SL, I pairs, $\omega = 1/2$), with the same additive, constant treatment effect, $\tau = 1/2$, in the favorable situation with $\{r_{Ci1} - r_{Ci2}\}/\omega \sim_{iid} F(.)$

Study	Distribution $F(\cdot)$	Number of Pairs	Dispersion ω	Power		
				$\Gamma = 1$	$\Gamma = 1.5$	$\Gamma = 2$
LM	Normal	120	1	1.00	0.96	0.60
SL	Normal	30	$\frac{1}{2}$	1.00	1.00	0.96
LM	Logistic	120	1	0.93	0.31	0.04
SL	Logistic	30	$\frac{1}{2}$	0.93	0.61	0.32
LM	Cauchy	200	1	0.98	0.32	0.02
SL	Cauchy	50	$\frac{1}{2}$	0.95	0.60	0.28

The power is similar when $\Gamma = 1$, but is higher for SL when Γ is larger

less heterogeneous study (SL) as

$$\widetilde{\Gamma} = \frac{\Phi\left(\sqrt{2}\frac{\tau}{\omega}\right)}{1 - \Phi\left(\sqrt{2}\frac{\tau}{\omega}\right)} \tag{16.1}$$

for Normal errors and as

$$\widetilde{\Gamma} = \frac{\Upsilon\left(\frac{\tau}{\omega}\right)}{1 - \Upsilon\left(\frac{\tau}{\omega}\right)} \tag{16.2}$$

for Cauchy errors. For Normal errors with $\tau = 1/2$, the designs LM ($\omega = 1$) and SL ($\omega = 1/2$) have design sensitivities $\widetilde{\Gamma}$ of 3.171 and 11.715, respectively. In light of this, the power of both designs with Normal errors tends to 1 as $I \to \infty$ for the values of Γ in Table 16.3, but for $\Gamma = 5$ the power of LM would tend to 0 while the power of SL would tend to 1. In parallel, for Cauchy errors with $\tau = 1/2$, the designs LM ($\omega = 1$) and SL ($\omega = 1/2$) have design sensitivities $\widetilde{\Gamma}$ of 1.838 and 3, respectively. In light of this, with Cauchy errors, for $\Gamma = 2$ in Table 16.3, the power of SL tends to 1 as $I \to \infty$ while the power of LM tends to 0.

16.3 Heterogeneity and the Sensitivity of Point Estimates

In the current chapter, the treatment has an additive constant effect, $\tau = r_{Tij} - r_{Cij}$, and in a randomized experiment, the Hodges–Lehmann estimate $\widehat{\tau}$ is a consistent estimate of τ. For a given deviation Γ from randomized treatment assignment, Tables 3.4 and 16.2 displayed the interval of possible Hodges–Lehmann point estimates $[\widehat{\tau}_{\min}, \widehat{\tau}_{\max}]$ of τ, where $\widehat{\tau}_{\min} = \widehat{\tau}_{\max} = \widehat{\tau}$ for $\Gamma = 1$; see Sect. 3.5 and [15]. As the sample size increases, $I \to \infty$, the endpoints of this interval converge in probability to the endpoints of a fixed interval, $[\tau_{\min}, \tau_{\max}]$; this interval reflects the uncertainty about τ that is due to a potential bias of magnitude Γ when there is no longer any sampling uncertainty.

In the "favorable situation," with errors having Normal $\Phi(\cdot)$ or Cauchy $\Upsilon(\cdot)$ cumulative distributions, the following proposition gives the form of this limiting interval. See [18, Appendix] for proof of Proposition 16.1.[2]

[2]Although the proof has a few details, it is simple in concept. The lower endpoint $\widehat{\tau}_{\min}$ of the interval of possible point estimates is obtained by equating Wilcoxon's signed rank statistic T computed from $Y_i - \tau_0$, say T_{τ_0}, to the maximum null expectation of T, namely $E\left(\overline{\overline{T}} \middle| \mathscr{F}, \mathscr{Z}\right) = \Gamma I(I+1)/\{2(1+\Gamma)\}$ from (3.19) and solving the equation for $\widehat{\tau}_{\min}$. Dividing the equation by $I(I+1)/2$ yields the equivalent equation $2T_{\tau_0}/\{I(I+1)\} = \Gamma/(1+\Gamma)$. If $(Y_i - \tau)/\omega \sim \Phi(\cdot)$ or $(Y_i - \tau)/\omega \sim \Upsilon(\cdot)$ then as $I \to \infty$, the left side of the equation, $2T_{\tau_0}/\{I(I+1)\}$, converges in probability to a function of $(\tau - \tau_0)/\omega$, and the rest of the proof is detail devoted to showing that equation can be solved to give the solutions in the statement of Proposition 16.1.

Proposition 16.1 *If* $(D_i - \tau)/\omega \sim_{iid} \Phi(\cdot)$ *then* $[\tau_{\min}, \tau_{\max}]$ *is* $\tau \pm \omega \Phi^{-1}(\kappa)/\sqrt{2}$, *where* $\kappa = \Gamma/(1+\Gamma)$. *If* $(D_i - \tau)/\omega \sim_{iid} \Upsilon(\cdot)$ *then* $[\tau_{\min}, \tau_{\max}]$ *is* $\tau \pm \omega \Upsilon^{-1}(\kappa)$.

Proposition 16.1 is consistent with Mill's view that heterogeneity of experimental units, ω, is directly relevant to causal claims. In Proposition 16.1, there is no sampling variability, because Proposition 16.1 refers to the limit as $I \to \infty$. The uncertainty addressed in Proposition 16.1 is quantified by the length of the limiting interval, $\tau_{\max} - \tau_{\min}$, and despite the absence of sampling variability, the length of that interval is directly proportional to ω.

Proposition 16.1 does not contradict Fisher's view, but it does emphasize that this view is applicable only when biases from nonrandom treatment assignment have been avoided by randomization. Reducing heterogeneity ω and increasing sample size I compete for resources in a randomized experiment because bias is known to have been avoided, so the analysis can be conducted with $\Gamma = 1$. More precisely, in Proposition 16.1, if it were known that $\Gamma = 1$, then $\kappa = 1/2$, so $\Phi^{-1}(\kappa)/\sqrt{2} = \Phi^{-1}(1/2)/\sqrt{2} = 0$ and $\Upsilon^{-1}(\kappa) = \Upsilon^{-1}(1/2) = 0$, and the length of the limiting interval is $\tau_{\max} - \tau_{\min} = 0$ for every value of ω.

The length of the interval is also affected by the magnitude of potential bias, Γ, though $\kappa = \Gamma/(1+\Gamma)$. The two components determine the length $\tau_{\max} - \tau_{\min}$ of the limiting interval in a multiplicative manner; for the Normal, $\tau_{\max} - \tau_{\min} = 2\omega\Phi^{-1}(\kappa)/\sqrt{2}$. A given magnitude Γ of deviation from a randomized experiment does more harm when the units are more heterogeneous, that is, when ω is larger. If you were deceptively trying to bias a randomized trial by covertly tilting the treatment assignment probabilities by a magnitude of Γ in (3.16)–(3.18), then you could do more harm if the units were the heterogeneous patients in a clinical trial than if they were the homogeneous genetically engineered mice in a laboratory experiment.

In the "favorable situation" in an observational study, increasing the sample size I reduces the standard error, but it does not materially reduce sensitivity to unmeasured biases. In contrast, in this situation, reducing the heterogeneity of experimental units, ω, reduces both the standard error and sensitivity to unmeasured biases. In an observational study, LM and SL of Fig. 16.1 are not at all the same: SL is much better.

16.4 Examples of Efforts to Reduce Heterogeneity

16.4.1 Twins

What are the economic returns to additional education? You cannot compare the mid-life earnings of surgeons and high school dropouts—they differed in the middle of high school before the dropout left school. Not in every case, but typically, the child who went on to become a surgeon was receiving better grades and standardized

test scores in high school, was more strongly motivated for school studies, had better educated, wealthier parents, and not inconceivably differed in some relevant genes. You would like to compare two children of the same parents with different education growing up at the same time in the same home with the same genes. Ashenfelter and Rouse [1] compared the earnings of identical twins with differing education, estimating about a 9% increase in earnings per year of additional education.

The use of twins is the canonical example of trading sample size for reduced heterogeneity. Twin pairs are quite heterogeneous between pairs and in several important respects fairly homogeneous within pairs, so the use of twins reflects the type of heterogeneity discussed in this chapter.

16.4.2 Road Hazards

What permanent features of a road affect the risk of collisions with roadside objects? Road hazards are a fairly small part of accident risk. Also relevant are: the driver's sobriety, skill, and risk tolerance; ambient light; the weather—ice, snow, rain, and fog; safety equipment—use of seat belts, quality and condition of brakes, tires, air bags, traction, and stability control devices. These factors are related. The risk-averse driver will drive near the legal speed limit, but will also invest in safety devices and wear seat belts. In the rain or snow, one drives on the highway to work but not on the dirt road to the picnic area, so weather and road hazards vary together. Sobriety is more common at noon than at midnight, so sobriety and ambient light vary together. You would like to compare different road hazards with the same driver, in the same car, in the same weather, with the same ambient light, in the same state of sobriety, with seat belts in the same state of use. Is this possible?

Using a simple, clever study design, Wright and Robertson [24] did just that. They compared 300 fatal roadside collisions in Georgia in 1974–1975 to 300 nonaccidents involving the same driver, in the same car, in the same ambient light, and so on. The nonaccidents occurred one mile back from the crash site, a location passed without incident by the driver just moments before the crash. At crash sites, Wright and Robertson found a substantial excess of roads that curved more than 6 degrees with downhill gradients of greater than 2%. (Technically, this is a "case-crossover" study of the type proposed by Malcolm Maclure [12], except that it is defined by geography rather than time; see also the "case-specular" design of Sander Greenland [4].)

16.4.3 The Genetically Engineered Mice of Microeconomics

Many businesses that provide products or services over large regions adopt a strategy known as "replication" in which nearly identical outlets are reproduced at high speed in diverse locations [23]. Starbucks and Tesco are two of the many such

businesses. This strategy confers various benefits to businesses that use it, but it also creates nearly identical copies of a business in locations that may have adopted different regulations, taxes, or other policies. For instance, Card and Kreuger's [2] study of the minimum wage and employment compared Burger Kings in New Jersey to Burger Kings in Pennsylvania, KFCs in New Jersey to KFCs in Pennsylvania, etc., and in this way eliminated one of several sources of extraneous variation between the two states; see Sects. 4.5 and 12.3 for further discussion of this study.

16.4.4 Motorcycle Helmets

Do helmets reduce the risk of death in motorcycle crashes? Crashes occur at different speeds with different forces, and neither speeds nor forces are likely to be measured. Motorcyclists hit different objects—pedestrians or Hummers—in dense or light traffic, with emergency services near or far away. One would like to compare two people, one with a helmet, the other without, on the same type of motorcycle, driving at the same speed, crashing into the same object, in the same traffic, with equal proximity to medical aid. Is that possible?

It is when two people ride one motorcycle, one with a helmet, the other without. Norvell and Cummings [14] looked at such crashes, finding about 40% lower risk associated with helmet use.

16.5 Summary

In a randomized experiment, an unbiased estimate of treatment effect is available, so increasing the sample size, I, or reducing the unit heterogeneity, ω, both serve to reduce the standard error of an unbiased estimate. The situation is strikingly different in an observational study. In the "favorable situation" in an observational study, the treatment is effective and there are no unmeasured biases. If the favorable situation arose, the investigator would not know it, and at best would hope to report that the treatment appears to be effective and that appearance is insensitive to small and moderate biases. In this situation, reducing heterogeneity, even purely random heterogeneity, ω, confers benefits that cannot be obtained by increasing the sample size, I. Specifically, reducing heterogeneity reduces sensitivity to unmeasured biases. Several cleverly designed studies have illustrated efforts to reduce heterogeneity.

16.6 Further Reading

This chapter is based on [18], where additional discussion may be found.

Various tactics in design and analysis can affect the heterogeneity of Y_i, and consequently the design sensitivity. At a basic level, we should pair closely for covariates that predict Y_i, and perhaps we can merely balance other covariates. For instance, if the presence of diabetes predicts surgical outcomes, but the region of the US does not, then pairs matched for the presence of diabetes are needed, and it may suffice to balance region of the US. Matching can be structured to reduce heterogeneity, thereby increasing insensitivity to unmeasured bias; see Sect. 11.8 and the paper by José Zubizarreta and colleagues [25].

Sometimes, the expectation or variance of Y_i varies as a function of matched observed covariates. Jesse Hsu and colleagues [9] propose a combination of exploratory analysis of $|Y_i|$ and confirmatory analysis of Y_i that is intended to locate subgroups of matched pairs with either larger typical effects or reduced heterogeneity of effect, such that these subgroups exhibit greater insensitivity to unmeasured biases. These subgroups are defined by values of observed covariates controlled by matching. The method is applied by Kwonsang Lee and colleagues [10].

Bibliography

1. Ashenfelter, O., Rouse, C.: Income, schooling and ability: evidence from a new sample of identical twins. Q. J. Econ. **113**, 253–284 (1998)
2. Card, D., Krueger, A.: Minimum wages and employment: a case study of the fast-food industry in New Jersey and Pennsylvania. Am. Econ. Rev. **84**, 772–793 (1994)
3. Fisher, R.A.: Design of Experiments. Oliver and Boyd, Edinburgh (1935)
4. Greenland, S.: A unified approach to the analysis of case-distribution (case-only) studies. Stat. Med. **18**, 1–15 (1999)
5. Hammond, E.C.: Smoking in relation to mortality and morbidity. J. Natl. Cancer Inst. **32**, 1161–1188 (1964)
6. Heller, R., Rosenbaum, P.R., Small, D.: Split samples and design sensitivity in observational studies. J. Am. Stat. Assoc. **104**, 1090–1101 (2009)
7. Hodges, J.L., Lehmann, E.L.: Estimates of location based on ranks. Ann. Math. Stat. **34**, 598–611 (1963)
8. Holland, P.W.: Statistics and causal inference. J. Am. Stat. Assoc. **81**, 945–960 (1986)
9. Hsu, J.Y., Zubizarreta, J.R., Small, D.S., Rosenbaum, P.R.: Strong control of the family-wise error rate in observational studies that discover effect modification by exploratory methods. Biometrika **102**, 767–782 (2015)
10. Lee, K., Small, D.S., Hsu, J.Y., Silber, J.H., Rosenbaum, P.R.: Discovering effect modification in an observational study of surgical mortality at hospitals with superior nursing. J. R. Stat. Soc. A **181**, 535–546 (2018)
11. Lehmann, E.L.: Nonparametrics, Holden Day, San Francisco (1975). Reprinted: Springer, New York (2006)
12. Maclure, M.: The case-crossover design: a method for studying transient effects on the risk of acute events. Am. J. Epidemiol. **133**, 144–152 (1991)

13. Mill, J.S.: A System of Logic: The Principles of Evidence and the Methods of Scientific Investigation. Liberty Fund, Indianapolis (1867)
14. Norvell, D.C., Cummings, P.: Association of helmet use with death in motorcycle crashes: a matched-pair cohort study. Am. J. Epidemiol. **156**, 483–487 (2002)
15. Rosenbaum, P.R.: Hodges-Lehmann point estimates of treatment effect in observational studies. J. Am. Stat. Assoc. **88**, 1250–1253 (1993)
16. Rosenbaum, P.R.: Observational Studies (2nd ed.). Springer, New York (2002)
17. Rosenbaum, P.R.: Design sensitivity in observational studies. Biometrika **91**, 153–164 (2004)
18. Rosenbaum, P.R.: Heterogeneity and causality: unit heterogeneity and design sensitivity in observational studies. Am. Stat. **59**, 147–152 (2005)
19. Rosenbaum, P.R.: What aspects of the design of an observational study affect its sensitivity to bias from covariates that were not observed? In: Festshrift for Paul W. Holland. ETS, Princeton (2009)
20. Salsburg, D.: The Lady Tasting Tea. Freeman, San Francisco (2001)
21. Small, D., Rosenbaum, P.R.: War and wages: the strength of instrumental variables and their sensitivity to unobserved biases. J. Am. Stat. Assoc. **103**, 924–933 (2008)
22. Werfel, U., Langen, V., Eickhoff, I., Schoonbrood, J., Vahrenholz, C., Brauksiepe, A., Popp, W., Norpoth, K.: Elevated DNA single-strand breakage frequencies in lymphocytes. Carcinogenesis **19**, 413–418 (1998)
23. Winter, S.G., Szulanski, G.: Replication as strategy. Organ. Sci. **12**, 730–743 (2001)
24. Wright, P.H., Robertson, L.S.: Priorities for roadside hazard modification. Traffic Eng. **46**, 24–30 (1976)
25. Zubizarreta, J.R., Parades, R.D., Rosenbaum, P.R.: Matching for balance, pairing for heterogeneity in an observational study of the effectiveness of for-profit and not-for-profit high schools in Chile. Ann. App. Stat. **8**, 204–231 (2014)

Chapter 17
Uncommon but Dramatic Responses to Treatment

Abstract Large effects in moderate to large studies are typically insensitive to small and moderate unobserved biases, but the concept of a "large effect" is vague. What if most subjects are not much affected by treatment, but a small fraction, perhaps 10% or 20% of subjects, are strongly affected? On average, such an effect may be small, but not at all small for the affected fraction. Is such an effect insensitive to small and moderate unobserved biases?

17.1 Large Effects, Now and Then

17.1.1 Are Large but Rare Effects Insensitive to Unmeasured Biases?

In Sect. 2.5, in the National Supported Work Experiment, depicted in Figure 2.1, many men appear to have received little or no benefit from the treatment, but the few men with high earnings were fairly consistently in the treated group. In Sect. 2.5, if pairs of men were examined two at a time (i.e., four men), then 61% of the time, the man with the highest aligned earnings was a treated man, where 50% is expected by chance, but if pairs were examined 20 at a time (i.e., 40 men), 86% of the time the man with the highest aligned earnings was a treated man, where again 50% is expected by chance. Big gains in earnings are consistently in the treated group.

A similar pattern is seen for toxicity in Chap. 8 in Fig. 8.1 in connection with Jeffrey Silber et al.'s [18] study of intensity of chemotherapy in $I = 344$ pairs of women with ovarian cancer: the median level of toxicity is not very different in the MO and GO groups, but high levels are more common in the MO group. If pairs of women are examined two at a time (i.e., 4 women), then in 65% of pairs, the woman with the highest aligned toxicity was treated by an MO, but if we look at 20 pairs at a time (i.e., 40 women), then 90% of the time, the woman with the highest aligned toxicity was treated by an MO. Figure 17.1 depicts the distribution of the MO-minus-GO differences, Y_i, in toxicity. In Figs. 8.1 and 17.1, it appears that many MOs and GOs treat with similar intensity producing similar toxicity, but a fraction of MOs treat more intensively producing more toxicity.

© Springer Nature Switzerland AG 2020

P. R. Rosenbaum, *Design of Observational Studies*, Springer Series in Statistics,
https://doi.org/10.1007/978-3-030-46405-9_17

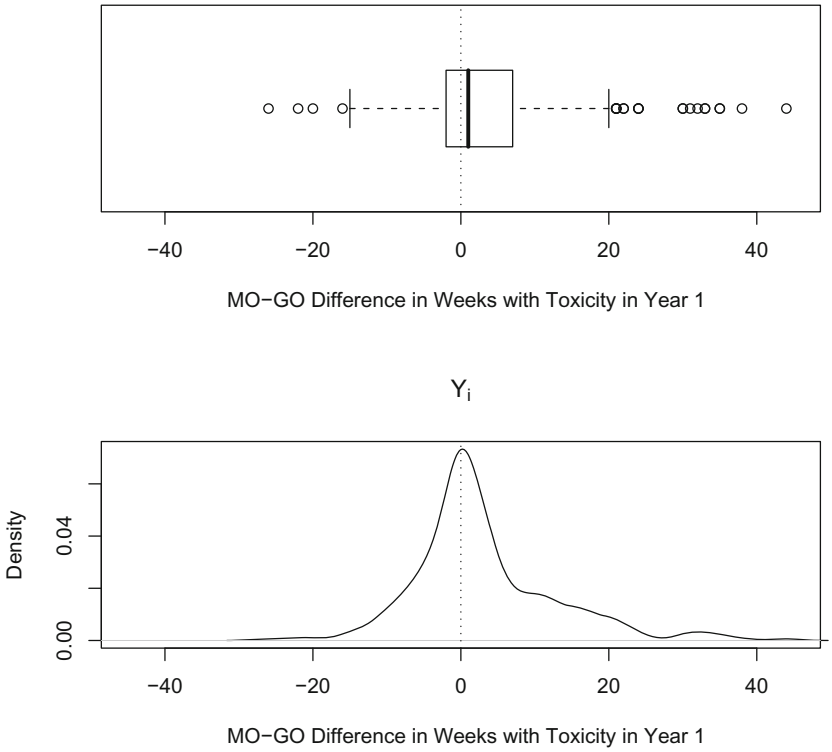

Fig. 17.1 Boxplot and density estimate of the MO-GO difference Y_i in weeks with toxicity in year 1 for $I = 344$ pairs of women with ovarian cancer; see Chap. 8 and [18]. The vertical dotted line is at zero difference. The density estimate uses the default settings in R

In these cases, the hypothesis of a constant effect, $\tau = r_{Tij} - r_{Cij}$ for all i, j, does not look plausible. As discussed in Sect. 2.5, a more plausible hypothesis is that $r_{Tij} - r_{Cij}$ is zero or small for many subjects i, j, but $r_{Tij} - r_{Cij}$ is large for some subjects i, j. David Salsburg [15] argued that effects of this sort are fairly common, are often important, and that we tend to miss them because our methodology tends to focus on typical effects, but in this context large effects do occur but are not typical. The question addressed in the current chapter is whether an effect of this kind is highly sensitive to unmeasured biases. In both Figs. 2.1 and 8.1, the typical difference in outcomes is not large, yet the distributions of outcomes are quite different.

17.1.2 Review of Sect. 2.5: Measuring Large but Uncommon Effects

Building upon certain technical results by Eric Lehmann [3, (6.1)], David Salsburg [15] and William Conover and Salsburg [1] made a strong argument that, in the comparison of unmatched treated and control groups, the pattern of large, uncommon effects is not properly associated with "transformations" or "outliers" or other similar concepts of data analysis, because the pattern may be seen in the ranks: the extremely high ranks are consistently in the treated group. In outline in Sect. 2.5 and in detail in [8], it was found that the highly interpretable rank scores, $\widetilde{q}_i$ in (2.7), proposed by W. Robert Stephenson [19] approximate the locally optimal rank scores of Conover and Salsburg [1]. For example, the statement about 20 pairs of men in the first paragraph of this Chapter was obtained in Sect. 2.5 by an interpretation of Stephenson's ranks.

Recall that Stephenson's ranks score pairs m at a time, and are essentially the same as Wilcoxon's ranks for $m = 2$; see Sect. 2.5. For $I = 344$ matched pairs, as in Silber et al.'s [13, 18] study in Chap. 8, Fig. 17.2 depicts Stephenson's ranks, $\widetilde{q}_i$ in (2.7), sorted into increasing order and scaled to be between 0 and 1 as $\widetilde{q}_i / \max \widetilde{q}_j$, for $m = 2, 5, 10, 20$. With $m = 20$, one looks at 20 pairs simultaneously, counting a "success" if the largest of the 40 aligned responses in the 20 pairs is the response of a treated subject. For $m = 10$ in Fig. 17.2, Stephenson's $\widetilde{q}_i$ largely ignore the smallest 200 of the 344 absolute differences in responses $|Y_i|$, judging the treatment effect by the remaining 144 pairs. Salsburg and Conover suggested the equivalent of $m = 5$ for the applications that they discussed.

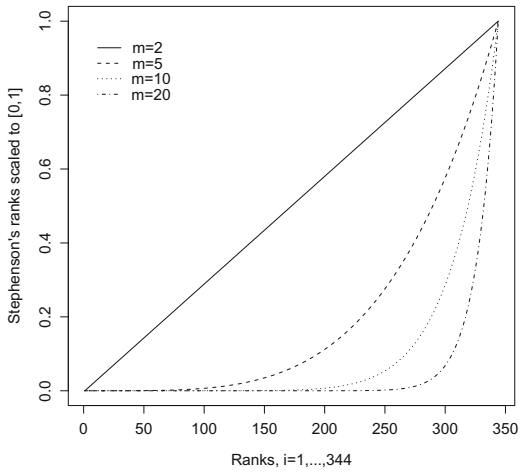

Fig. 17.2 Stephenson's ranks $\widetilde{q}_i$ from (2.7) for $m = 2, 5, 10, 20$ and for the $I = 344$ matched pairs in Silber et al.'s [18] study of intensity of chemotherapy for ovarian cancer; see Chap. 8. Wilcoxon's ranks are essentially the same as $m = 2$. For $m = 5$, the smallest 100 or so $|Y_i|$ are almost ignored. For $m = 20$, the smallest 300 or so $|Y_i|$ are almost ignored

17.2 Two Examples

17.2.1 Chemotherapy Intensity and Toxicity in Ovarian Cancer

Table 17.1 compares four sensitivity analyses for the ovarian cancer study [13, 18] in Chap. 8. The four sensitivity analyses are each conducted using Stephenson's [19] signed rank test, with four values of m, that is, with the absolute ranks in Fig. 17.2. Wilcoxon's ranks are similar to Stephenson's ranks with $m = 2$. The ranks with $m = 20$ are concerned with the largest aligned response of 40 women in 20 pairs, so that they emphasize the fairly high levels of toxicity experienced by some women.

Table 17.1 reports upper bounds on the one-sided P-value testing the null hypothesis of no treatment effect. Wilcoxon's statistic ($m = 2$) becomes sensitive to a bias of magnitude $\Gamma = 1.5$, but Stephenson's statistic with $m = 20$ is insensitive to a bias of magnitude $\Gamma = 3.5$. In this one example, the degree of sensitivity to bias is strongly affected by whether the analysis emphasizes typical effects or less common but larger effects. Small biases could not produce the large but uncommon effects in Fig. 17.1.

The calculations to produce Table 17.1 closely parallel the calculations for Wilcoxon's signed rank statistic, except that Stephenson's ranks, $\widetilde{q}_i$, are used in place of Wilcoxon's ranks, q_i. The null expectation and variance of Stephenson's statistic $\widetilde{T} = \sum \text{sgn}\,(Y_i)\ \widetilde{q}_i$ are bounded by expressions (3.25)–(3.26) with $\widetilde{q}_i$ in place of q_i, and the standardized deviates are compared to the Normal distribution in parallel with (3.23).

Table 17.1 Using Stephenson's statistic with several values of m, upper bounds are given on the one-sided P-value testing the null hypothesis of no treatment effect in the ovarian cancer data of Chap. 8 and [18]

Γ	$m = 2$	$m = 5$	$m = 10$	$m = 20$
1	9.0×10^{-7}	5.2×10^{-9}	1.1×10^{-8}	1.1×10^{-6}
1.5	0.051	0.00017	0.000024	0.00019
2	0.71	0.016	0.0010	0.0020
2.5		0.14	0.0087	0.0081
3			0.035	0.021
3.5			0.088	0.041

For $m = 2$, Stephenson's ranks closely resemble Wilcoxon's ranks, but for larger values of m, Stephenson's ranks place greater emphasis on large $|Y_i|$'s. In this instance, the results are much less sensitive with $m = 20$ than with $m = 2$

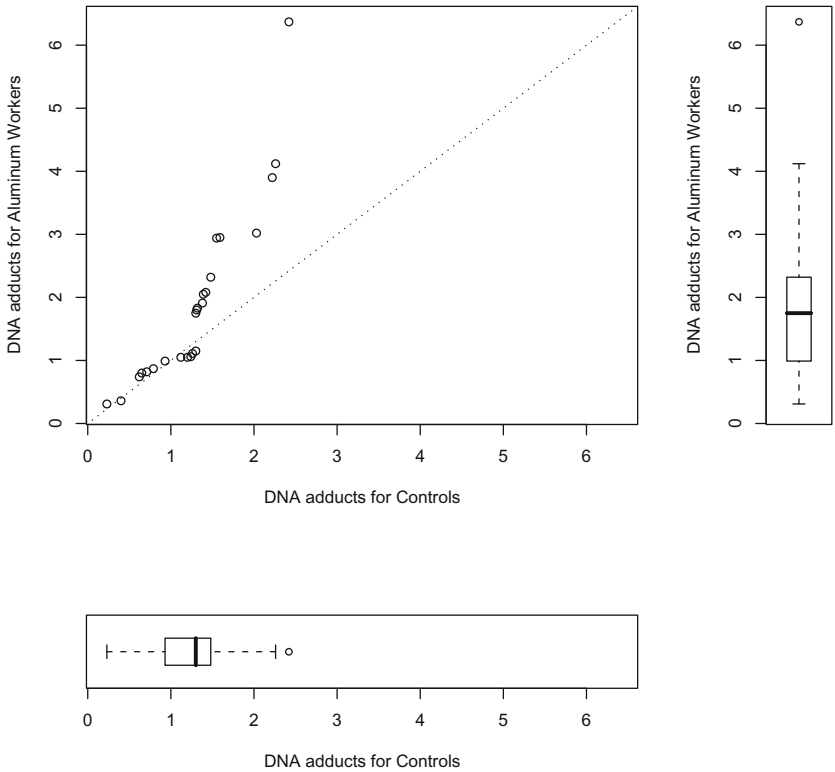

Fig. 17.3 DNA adducts per 10^8 nucleotides for 25 aluminum production workers and 25 controls matched for age and cigarettes per day [17]. The central plot is a quantile–quantile plot that ignores the pairing. The dotted line is the $x = y$ line. There appears to be little or no difference at lower quantiles, with a sharp rise at upper quantiles

17.2.2 DNA Adducts Among Aluminum Production Workers

Bernadette Schoket, David Phillips, Alan Hewer, and István Vincze [17] compared DNA adducts in the lymphocytes of aluminum production workers and unexposed controls. Figure 17.3 shows 25 matched pairs of an aluminum production worker and a control matched for age and cigarettes smoked per day. The noticeable feature of Fig. 17.3 is that there appears to be little or no difference in DNA adducts at lower quantiles, with a substantial divergence at upper quantiles.[1]

Table 17.2 displays two sensitivity analyses using Stephenson's signed rank statistic, one with $m = 2$ which is similar to Wilcoxon's statistic, and one with $m = 5$. (With just $I = 25$ pairs, it does not seem reasonable to compare $m = 20$

[1]The paired data for the 25 pairs are listed in [8].

Table 17.2 Upper bounds on the one-sided P-values from Stephenson's test for $m = 2$ and $m = 5$ for the 25 pairs of an aluminum production worker and a matched control depicted in Fig. 17.3

Γ	$m = 2$	$m = 5$
1	0.0076	0.0078
1.5	0.054	0.030
1.8	0.10	0.048
2	0.14	0.061

The results are less sensitive to bias for $m = 5$, presumably because the difference in the distribution of DNA adducts is apparent only at upper quantiles

pairs at a time to determine the largest aligned response.) The table displays upper bounds on the one-sided P-value for testing the null hypothesis of no treatment effect. As Fig. 17.3 might lead us to anticipate, the results are less sensitive for $m = 5$ than for $m = 2$.

17.3 Properties of a Paired Version of Salsburg's Model

David Salsburg [15] proposed a model for treatment effects in unmatched groups in which a fraction $0 \leq \lambda \leq 1$ of the treated subjects experience a substantial treatment effect and the rest experience no effect at all. The following model is a slight modification for matched pairs of Salsburg's model for unmatched groups. Each pair i is characterized by a fixed pair parameter, η_i, which accounts for dependence within pairs, and is eliminated when matched pair differences Y_i are calculated. Then the $r_{Cij} - \eta_i$ are sampled from a continuous distribution $G(\cdot)$ with density $g(\cdot)$, and the $r_{Tij} - \eta_i$ are sampled from $(1-\lambda)\, G(\cdot) + \lambda \{G(\cdot)\}^{\nu}$ where $0 \leq \lambda \leq 1$ and $\nu \geq 2$ is a positive integer. Responses from distinct subjects are mutually independent.[2]

The model has a simple interpretation. The model says, in effect, that a fraction, $1 - \lambda$, of treated subjects experience no treatment effect—their responses are sampled from $G(\cdot)$ which is the distribution of responses to control. A fraction

[2]The density of $r_{Tij} - \eta_i$ is $(1-\lambda)\ g(r) + \lambda\, \nu\, g(r)\ \{G(r)\}^{\nu-1}$. In the favorable situation, treatment assignment $Z_{i1} - Z_{i2}$ is ± 1 with probability $1/2$ independently of $(r_{Ci1}, r_{Ti1}, r_{Ci2}, r_{Ti2})$, and $Y_i = (Z_{i1} - Z_{i2})\,(R_{i1} - R_{i2})$ is distributed as the difference of two independent observations, one drawn from $(1-\lambda)\, G(\cdot) + \lambda \{G(\cdot)\}^{\nu}$ and the other drawn from $G(\cdot)$. The matched pair difference Y_i has probability density

$$\int g(r) \left[(1-\lambda)\ g(r+y) + \lambda\, \nu\, g(r+y)\ \{G(r+y)\}^{\nu-1} \right] dr$$

which is easy to evaluate by numerical integration. In the calculations presented graphically, $G(\cdot)$ is the standard Normal cumulative distribution $\Phi(\cdot)$.

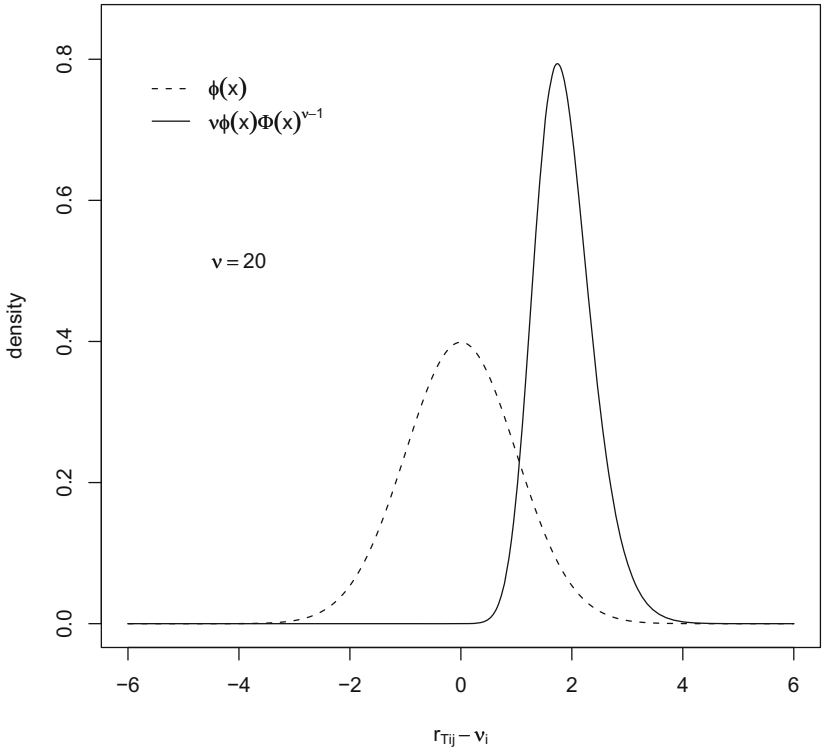

Fig. 17.4 The two densities that are mixed in Salsburg's model to form the distribution of $r_{Tij} - \eta_i$, with $\nu = 20$. The dashed density function is the standard Normal density. The solid density function is the density of the maximum of $\nu = 20$ independent observations from the standard Normal distribution

λ of treated subjects do experience an effect. Instead of having a response drawn from $G(\cdot)$, the affected fraction have a response $r_{Tij} - \eta_i$ drawn from $\{G(\cdot)\}^{\nu} = G(\cdot) \times \cdots \times G(\cdot)$. If you sampled ν independent $r_{Cij} - \eta_i$'s from the distribution $G(\cdot)$ and took the maximum of these, that maximum would have the distribution $\{G(\cdot)\}^{\nu}$. As intuition suggests, $\{G(\cdot)\}^{\nu}$ is a "larger" distribution[3] than $G(\cdot)$ in the sense that $\{G(r)\}^{\nu} \leq G(r)$ for all r, because $\nu \geq 2$; that is, the chance, namely $\{G(r)\}^{\nu}$, that the maximum of ν independent observations from $G(\cdot)$ is less than r is less than or equal to the chance, namely $G(r)$, that one observation from $G(\cdot)$ is less than r, and this is true for all r. Figure 17.4 depicts the two probability densities that are mixed with weights λ and $1 - \lambda$ to form the density of the treated response, $r_{Tij} - \eta_i$, for $\nu = 20$, when $G(\cdot)$ is the standard Normal cumulative distribution $G(\cdot) = \Phi(\cdot)$. In Fig. 17.4, a fraction $1 - \lambda$ of treated subjects have a response $r_{Tij} - \eta_i$ drawn from the standard Normal density $\phi(r)$, the dashed curve, and a

[3]It is stochastically larger; see Note 15 in Chap. 3.

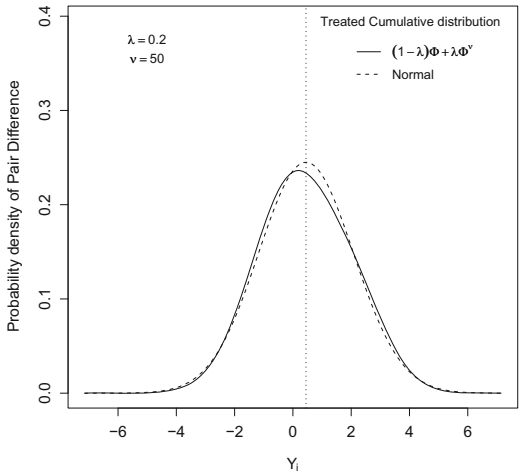

Fig. 17.5 Salsburg's model (solid line) for the matched pair difference Y_i with $\lambda = 0.2$ and $\nu = 50$ in the treated group. For comparison, a Normal density (dashed line) of Y_i with the same expectation and variance is plotted. The vertical dotted line is at their shared expectation

fraction λ of treated subjects have a response $r_{Tij} - \eta_i$ drawn from the density $\nu\,\phi(r)\,\Phi(r)^{\nu-1}$ of the maximum of $\nu = 20$ independent observations from a standard Normal distribution, the solid curve. The treated-minus-control difference Y_i in pair i is the difference of a treated response $r_{Tij} - \eta_i$ formed from this mixture and a control response $r_{Cij} - \eta_i$ independently drawn from the standard Normal density $\phi(r)$.

In the unmatched two sample problem, Eric Lehmann [3, (6.1)] studied this model for $\nu = 2$ showing that Wilcoxon's ranks, with $m = 2$, perform well for small λ in large samples. William Conover and David Salsburg [1] showed that for general ν, ranks similar to Stephenson's ranks perform well if $m = \nu$ for small λ in large samples.[4]

With $G(\cdot)$ equal to the standard Normal distribution, $G(\cdot) = \Phi(\cdot)$, Fig. 17.5 depicts the distribution of matched pair differences, Y_i, under Salsburg's model with $\lambda = 0.2$ and $\nu = 50$. For comparison, Fig. 17.5 also displays a Normal distribution with the same expectation and variance. Visually, the distinction between the two distributions is quite subtle: Salsburg's model has a slight skew to the right, with an expectation slightly to the right of its mode. Conceptually, the distributions are very different. In Salsburg's model, $1 - \lambda = 80\%$ of treated subjects experience no treatment effect: their treated responses, $r_{Tij} - \eta_i$, are sampled from the same

[4]Somewhat more precisely, Conover and Salsburg found the locally most powerful rank test for fixed ν as $\lambda \to 0$ for the two-sample problem as the sample size increased. It is shown in [8] that these locally most powerful ranks are almost the same as Stephenson's [19] ranks in large samples, although the latter are easier to interpret in terms of the magnitude of the treatment effect. Stephenson and Ghosh [20] discuss the use of Stephenson's ranks in the unmatched two sample problem. The two-sample model comparing $(1 - \lambda)\,G(\cdot) + \lambda\,\{G(\cdot)\}^{\nu}$ and $G(\cdot)$ is an instance of a "Lehmann alternative" in the sense that the distribution of the ranks depends upon ν and λ but not on $G(\cdot)$; see [3]. Reference [8] is not restricted to matched pairs, but in the current book, only the paired case is discussed.

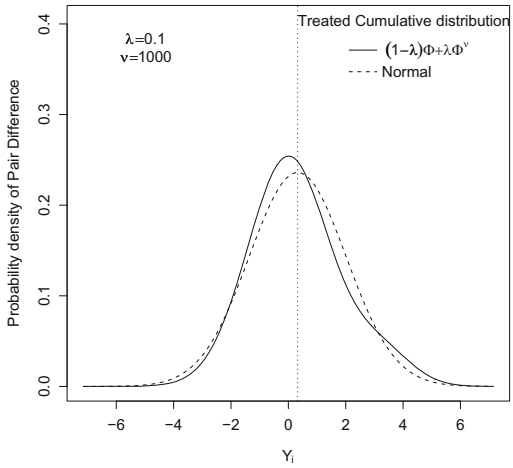

Fig. 17.6 Salsburg's model (solid line) for the matched pair difference Y_i with $\lambda = 0.1$ and $\nu = 1000$ in the treated group. For comparison, a Normal density (dashed line) of Y_i with the same expectation and variance is plotted. The vertical dotted line is at their shared expectation

Normal distribution as responses to control, $r_{Cij} - \eta_i$. In Salsburg's model, only $\lambda = 20\%$ of treated subjects experience a nonzero effect, and their responses are the maximum of $\nu = 20$ independent observations from a Normal distribution, rather than one observation from a Normal distribution. In Fig. 17.4, the affected group had a much higher distribution of responses, but this is twice hidden in Fig. 17.5, once because only a fraction λ of treated subjects experienced a treatment effect, and a second time by the need to compare treated subjects to controls to eliminate η_i by differencing to form the matched pair difference, Y_i. Figure 17.6 is similar to Fig. 17.5, except $\lambda = 0.1$ and $\nu = 1000$; here, the difference is more noticeable, but still quite subtle. Compare Fig. 17.6 to the estimated density in Fig. 17.1.

Are the effects in Figs. 17.5 and 17.6 highly sensitive to unmeasured biases? On the one hand, the affected group in Fig. 17.4 is strongly affected. On the other hand, that strong effect is twice hidden in the observable distribution of Y_i in Fig. 17.5. In Tables 17.1 and 17.2, effects were judged fairly sensitive by Wilcoxon's statistic ($m = 2$), which emphasizes typical effects, but were noticeably less sensitive to bias when Stephenson's statistic was used with larger values of m. Is the pattern in Tables 17.1 and 17.2 something that can be expected under Salsburg's model?

17.4 Design Sensitivity for Uncommon but Dramatic Effects

17.4.1 Design Sensitivity of Stephenson's Test

To answer this question, the design sensitivity $\widetilde{\Gamma}$ of Stephenson's [19] test is determined in general. Then $\widetilde{\Gamma}$ is calculated for Salsburg's model.

As in Chap. 15, the favorable situation is assumed, so there is a treatment effect without unmeasured biases, and therefore one hopes to report that rejection of

the null hypothesis of no treatment effect is insensitive to small and moderate unmeasured biases. Therefore, suppose that the Y_i are independent observations from some continuous distribution $H(\cdot)$ and there are no unmeasured biases, in the sense that treatments were assigned by the naïve model (3.5)–(3.8).

Consider m independent observations Y_i drawn from $H(\cdot)$. One of these m observations, say Y_ℓ, has the largest value of $|Y_i|$; that is, $|Y_\ell| = \max_{1 \le i \le m} |Y_i|$. Define $\widetilde{p}_m$ as the probability that the one Y_ℓ among these m with the largest value of $|Y_i|$ is positive, $Y_\ell > 0$.

The first thing to notice about $\widetilde{p}_m$ is that we have already made extensive use of $\widetilde{p}_2$ in Chap. 15 because $\widetilde{p}_2 = p_1^{'} = \Pr(Y_i + Y_{i'} > 0)$; that is, for $i \neq i'$, the sign of $Y_i + Y_{i'}$ is determined by Y_i if $|Y_i| > |Y_{i'}|$ and by $Y_{i'}$ if $|Y_i| < |Y_{i'}|$. Indeed, by Proposition 15.1, the design sensitivity $\widetilde{\Gamma}$ of Wilcoxon's statistic ($m = 2$) is $p_1^{'} / \left(1 - p_1^{'}\right) = \widetilde{p}_2 / (1 - \widetilde{p}_2)$. The second thing to notice about $\widetilde{p}_m$ is that it can be easily calculated by simulation for any distribution $H(\cdot)$.

The following proposition generalizes Proposition 15.1.

Proposition 17.1 *In I pairs matched exactly for observed covariates, $\mathbf{x}_{ij}$, suppose that the treated-minus-control matched pair differences in outcomes, Y_i, $i = 1, \ldots, I$, are independently sampled from a distribution $H(\cdot)$ and there is no bias from unmeasured covariates in the sense that treatments are actually assigned by the naïve model (3.5)–(3.8). A sensitivity analysis is performed using Stephenson's signed rank statistic $\widetilde{T}$ with a specific value of m and Γ. The power of this sensitivity analysis for one-tailed α-level test of no effect with $\alpha > 0$ tends to 1 as $I \to \infty$ for $\Gamma < \widetilde{\Gamma}$ and the power tends to 0 for $\Gamma > \widetilde{\Gamma}$ where*

$$\widetilde{\Gamma} = \frac{\widetilde{p}_m}{1 - \widetilde{p}_m}. \tag{17.1}$$

The proof of Proposition 17.1 is similar to the proof of Proposition 15.1 in Chap. 15 and is given in detail in [9] and is sketched in the appendix to this chapter.

17.4.2 Design Sensitivity of Stephenson's Test under Salsburg's Model

Table 17.3 reports design sensitivities $\widetilde{\Gamma}$ for Stephenson's [19] signed rank statistic when applied to the paired version of Salsburg's [15] model. In Table 17.3, either $\lambda = 10\%$ or $\lambda = 20\%$ of treated subjects are affected by the treatment, receiving either the maximum of $\nu = 50$ or $\nu = 1000$ independent responses from the control distribution $G(\cdot)$, with $G(\cdot)$ equal to either the Normal or Cauchy distributions. The densities of two of these distributions were shown in Figs. 17.5 and 17.6.

If only a small fraction of treated subjects respond to treatment, then the use of Wilcoxon's ranks ($m = 2$) is a mistake. In this case, much larger values of the design sensitivity $\widetilde{\Gamma}$ are obtained by using a test that focuses attention on large values of

Table 17.3 Design sensitivity of Stephenson's test based on subsets of m pairs for Salsburg's model in which only a fraction λ of treated subjects are affected by the treatment, having a response that is the maximum of ν observations from the distribution of responses to control, $G(.)$, for $G(.)$ equal to the Normal or Cauchy distributions

	Normal	Normal	Cauchy	Cauchy
	$\lambda = 0.2,\ \nu = 50$	$\lambda = 0.1,\ \nu = 1000$	$\lambda = 0.2,\ \nu = 50$	$\lambda = 0.1,\ \nu = 1000$
$m = 2$	1.8	1.4	2.0	1.5
$m = 5$	2.6	2.0	3.3	2.3
$m = 10$	3.4	3.0	4.7	4.4
$m = 20$	4.4	4.9	5.7	11.5

$|Y_i|$. This conclusion is specific to the situation in which many treated subjects are unaffected by treatment and a small fraction are strongly affected.

17.5 Summary

A treatment may affect some subjects and not others. In this case, the typical effect may be small, despite large effects for some subjects. Conventional statistical methods look for typical treatment effects and may miss dramatic effects for a small subset of affected subjects. The issue arises in both randomized clinical trials [1, 15, 16] and observational studies [8, 9]. When the treatment has an effect of this form in an observational study, an analysis focused on detecting such an effect, for instance, using Stephenson's [19] test, may be substantially less sensitive to unmeasured bias than an analysis focused on typical effects. The issue is relevant to planning statistical analyses as discussed in Part IV.

17.6 Further Reading

This chapter is built from components found in papers by Eric Lehmann [3, (6.1)], W. Robert Stephenson [19], David Salsburg [1, 15, 16], and William Conover [1]; see also [8, 9] for discussion of the use of these components in observational studies.

Stephenson's statistic is a member of a larger class of U-statistics built for superior design sensitivity. See Sect. 19.2 and [10, 11].

For reasons that are not completely understood, an extremely stressful event produces posttraumatic stress symptoms in some people, but not in others. Perhaps the "same" event has different physical or social consequences for people differently situated, or perhaps people exhibit unequal resilience in the face of stress. The methods of this chapter were applied in a study by José Zubizarreta and colleagues [22] of posttraumatic stress following the 2010 earthquake in Chile.

An alternative approach emphasizes the possibility that the magnitude of the treatment effect $r_{Tij} - r_{Cij}$ is not constant but rather increasing in r_{Cij}—a so-called dilated treatment effect—and estimates the effect at different quantiles of r_{Cij}; see [4, 6, §5]. With a dilated treatment effect, the effect at upper quantiles may be less sensitive to bias than the effect at lower quantiles.

17.7 Software

With appropriate settings, the function `senU` in the `R` package `DOS2` performs Stephenson's test and its sensitivity analysis. See also the related discussion of the U-statistic in Sect. 19.2. Stephenson's statistic is a special case of the U-statistic implemented by `senU`.

The sensitivity analysis for a dilated effect in Sect. 17.6 entails cutting a continuous outcome at an upper quantile of the pooled distribution of outcomes in treated and control groups, thereby producing a binary outcome; see [4] for the easy steps. Sensitivity analysis for the resulting binary outcome is implemented in the `mh` function of the `sensitivity2x2xk` package. .

17.8 Data

The data of Schoket et al. [17] in Fig. 17.3 are in the object `schoket` in the `R` package `DOS2`. The examples in the documentation for `schoket` reproduce and expand upon Table 17.2.

Appendix: Sketch of the Proof of Proposition 17.1

As in the proof of Proposition 15.1 in Chap. 15, there are two versions of the proof. One version uses the heuristic and appeals to the general fact from [7, §3] that the heuristic works quite generally. The other version is complete as it stands and parallels the proof in Chap. 15. In either case, one starts with the observation from Stephenson [19] that in the favorable situation a scaled version of $\widetilde{T}$ converges in probability to $\widetilde{p}_m$, specifically

$$\frac{m!(I-m)!}{I!} \quad \widetilde{T} \quad \rightarrow \quad \widetilde{p}_m.$$

One then checks that under the sensitivity analysis model, the maximum null expectation of $\widetilde{T}$ after scaling in the same way equals $\Gamma/(1+\Gamma)$. Equating $\widetilde{p}_m = \Gamma/(1+\Gamma)$ and solving completes the heuristic derivation. If no appeal is made

to [7, §3], the detailed proof must again verify that if $\widetilde{p}_m \neq \Gamma/(1+\Gamma)$, then the difference in expectations overwhelms the variance terms as $I \to \infty$.

Bibliography

1. Conover, W.J., Salsburg, D.S.: Locally most powerful tests for detecting treatment effects when only a subset of patients can be expected to 'respond' to treatment. Biometrics **44**, 189–196 (1988)
2. Koenker, R.: Rank tests for heterogeneous treatment effects with covariates. In: Nonparametrics and Robustness, pp. 134–142. IMS, Hayward (2010)
3. Lehmann, E.L.: The power of rank tests. Ann. Math. Stat. **24**, 23–43 (1953)
4. Rosenbaum, P.R.: Reduced sensitivity to hidden bias at upper quantiles in observational studies with dilated treatment effects. Biometrics **55**, 560–564 (1999)
5. Rosenbaum, P.R.: Effects attributable to treatment: inference in experiments and observational studies with a discrete pivot. Biometrika **88**, 219–231 (2001)
6. Rosenbaum, P.R.: Observational Studies (2nd ed.). Springer, New York (2002)
7. Rosenbaum, P.R.: Design sensitivity in observational studies. Biometrika **91**, 153–164 (2004)
8. Rosenbaum, P.R.: Confidence intervals for uncommon but dramatic responses to treatment. Biometrics **63**, 1164–1171 (2007)
9. Rosenbaum, P.R.: What aspects of the design of an observational study affect its sensitivity to bias from covariates that were not observed? In: Dorans, N.J., Sinharay, S. (eds.) Looking Back: Proceedings of a Conference in Honor of Paul W. Holland. Lecture Notes in Statistics Book 202, pp. 87–114. Springer, New York (2011)
10. Rosenbaum, P.R.: A new U-statistic with superior design sensitivity in matched observational studies. Biometrics **67**, 1017–1027 (2011)
11. Rosenbaum, P.R.: Bahadur efficiency of sensitivity analyses in observational studies. J. Am. Stat. Assoc. **110**, 205–217 (2015)
12. Rosenbaum, P.R., Silber, J.H.: Amplification of sensitivity analysis in observational studies. J. Am. Stat. Assoc. **104**, 1398–1405 (2009)
13. Rosenbaum, P.R., Ross R.N., Silber, J.H.: Minimum distance matched sampling with fine balance in an observational study of treatment for ovarian cancer. J. Am. Stat. Assoc. **102**, 75–83 (2007)
14. Rosenblatt, J.D., Benjamini, Y.: On mixture alternatives and Wilcoxon's signed-rank test. Am. Stat. **72**, 344–347 (2018)
15. Salsburg, D.S.: Alternative hypotheses for the effects of drugs in small-scale clinical studies. Biometrics **42**, 671–674 (1986)
16. Salsburg, D.S.: The Use of Restricted Significance Tests in Clinical Trials. Springer, New York (1992)
17. Schoket, B., Phillips, D.H. , Hewer, A., Vincze, I.: ^{32}P-Postlabelling detection of aromatic DNA adducts in peripheral blood lymphocytes from aluminum production plant workers. Mutat. Res. **260**, 89–98 (1991)
18. Silber, J.H., Rosenbaum, P.R., Polsky, D., Ross, R.N., Even-Shoshan, O., Schwartz, S., Armstrong, K.A., Randall, T.C. : Does ovarian cancer treatment and survival differ by the specialty providing chemotherapy? J. Clin. Oncol. **25**, 1169–1175 (2007)
19. Stephenson, W.R.: A general class of one-sample nonparametric test statistics based on subsamples. J. Am. Stat. Assoc. **76**, 960–966 (1981)
20. Stephenson, W.R., Ghosh, M. : Two sample nonparametric tests based on subsamples. Commun. Stat. **14**, 1669–1684 (1985)

21. Wilcoxon, F.: Individual comparisons by ranking methods. Biometrics **1**, 80–83 (1945)
22. Zubizarreta, J.R., Cerda, M., Rosenbaum, P.R.: Effect of the 2010 Chilean earthquake on posttraumatic stress: reducing sensitivity to unmeasured bias through study design. Epidemiol **7**, 79–87 (2013)

Chapter 18
Anticipated and Discovered Patterns of Response

Abstract Design sensitivity is used to quantify the effectiveness of devices discussed in Chap. 5. Several of those devices anticipate a particular pattern of results, perhaps coherence among several outcomes, or a dose–response relationship. To what extent do these considerations reduce sensitivity to unmeasured biases? Must a coherent pattern of associations be anticipated? Or is it possible to discover a coherent pattern in the data at hand with suitable corrections for multiple testing?

18.1 Using Design Sensitivity to Evaluate Devices

In Sect. 5.2, several devices used to study unmeasured biases made reference to an anticipated pattern of responses. In Sect. 5.2.3, it was anticipated that several outcomes would be affected in a similar way, the pattern of coherence. In Sect. 5.2.5, it was anticipated that larger doses of treatment would yield larger responses, the pattern of dose-response. If such a pattern is anticipated, and if the anticipation is realized in the observed data, will the sensitivity to unmeasured biases be reduced? This did occur in particular examples in Tables 5.3 and 5.5. Under what circumstance should similar results be expected?

Must a coherent pattern of associations be anticipated if it is to affect design sensitivity? Or is it possible to discover coherence in the data at hand? Suppose that multiple tests are employed to discover coherence, but suitable corrections are made for testing multiple hypotheses. Can the discovery of coherence in data reduce sensitivity to unmeasured biases? Is the design sensitivity of discovered coherence inferior to the design sensitivity of anticipated coherence?

© Springer Nature Switzerland AG 2020

P. R. Rosenbaum, *Design of Observational Studies*, Springer Series in Statistics,
https://doi.org/10.1007/978-3-030-46405-9_18

18.2 Coherence

18.2.1 Notation with Several Responses

In Sect. 5.2.3, the coherent signed rank statistic ([12] and [13, §9]) was simply the sum of the several signed rank statistics for several outcomes oriented in the anticipated direction. For instance, in Table 5.3, the signed rank statistics for the Math and Verbal test scores were added together, creating a statistic that would react strongly to improvements in both Math and Verbal scores, and weakly to, say, an increase in Math scores together with a decline in Verbal scores.

The coherent signed rank statistic so defined has advantages in analysis: the separate ranking of each outcome is one robust method for standardizing the outcomes, giving about equal weight to each of the outcomes. In studying design sensitivity for coherence, it is easiest if a slightly different statistic is used: Wilcoxon's signed rank statistic is applied to a weighted combination of the outcomes. In the first approach, weights are applied to the separate signed rank statistics, while in the second approach a single signed rank statistic is computed from a weighted combination of the outcomes. If the second approach is adopted, then in the multivariate Normal situation, there is an explicit formula for the design sensitivity for given weights, an explicit formula for the optimal weights, and so on. In analysis, data are not typically multivariate Normal, and the second approach leaves the investigator with the task of finding an appropriate, robust standardization of the outcomes. It is possible to compute the design sensitivity for both approaches, but the first approach requires some unenlightening Monte Carlo work; however, when this is done, similar results are found for both approaches. Ruth Heller, Dylan Small, and I [4] looked at both approaches in quite a bit of detail; see also [15, 20]. Here, a few of the simpler results from [4] are presented, and the reader may turn to [4] and [15] for additional approaches and numerical results.[1]

Instead of a single observed response R_{ij} from the jth person in pair i, that person now has M observed responses, $\mathbf{R}_{ij} = \left(R_{ij1}, \ldots, R_{ijM}\right)^T$. In Table 5.3, $M = 2$, and R_{ij1} was the Math score, while R_{ij2} was the Verbal score. Wilcoxon's signed rank statistic will be applied to a weighted combination of the responses, $\sum_{m=1}^{M} \lambda_m R_{ijm}$, where the weights λ_m are selected before the data are collected. A data-dependent choice of λ_m is discussed in Sect. 18.3. Write $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_M)^T$

[1] Why are simpler results available if the signed rank statistic is applied to a weighted combination of results, rather than applying weights to several signed rank statistics? As discussed in Sect. 2.3.3, the signed rank statistic is distribution-free; that is, its null distribution can be determined before any data are collected. In contrast, the coherent signed rank statistic is a randomization test in a randomized experiment, but, like the sample mean in Chap. 2, it is not distribution-free: its randomization distribution depends upon features of the observed data. When thinking about design, with no data on hand, it is convenient that the null distribution of the test statistic is known and does not depend upon the data you do not have. On the other hand, in principle, the design sensitivity is well defined in either case, but more Monte Carlo work may be needed to determine its value. See [4] for some comparisons.

for the column vector with M coordinates λ_m. More precisely, Wilcoxon's signed rank statistic, T, is computed from the treated-minus-control difference in combined responses,

$$Y_i = (Z_{i1} - Z_{i2}) \left(\sum_{m=1}^{M} \lambda_m \, R_{i1m} - \sum_{m=1}^{M} \lambda_m \, R_{i2m} \right) \tag{18.1}$$

$$= \sum_{m=1}^{M} \lambda_m \, Y_{im}, \tag{18.2}$$

where

$$Y_{im} = (Z_{i1} - Z_{i2}) \, (R_{i1m} - R_{i2m}) \, . \tag{18.3}$$

Again, this differs slightly from Table 5.3 where weights $\lambda_1 = 1$ and $\lambda_2 = 1$ were applied to two signed rank statistics, one computed from R_{ij1}, the other from R_{ij2}.

18.2.2 Responses with a Multivariate Normal Distribution

As in Chap. 15, design sensitivity is computed in the "favorable situation" in which there is a treatment effect and no bias from unmeasured covariates. It is in the favorable situation that we hope to report that rejection of the hypothesis of no effect is insensitive to unmeasured biases.

The results are particularly simple for multivariate Normal $\mathbf{R}_{ij}$.[2] The multivariate Normal distribution may be defined in various equivalent ways, but one definition is that $\sum_{m=1}^{M} \lambda_m \, R_{ijm}$ has some Normal distribution for every choice of weights, $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_M)^T$. In particular, it is assumed that each response R_{ijm} is affected by an additive treatment effect τ_m, so $R_{ijm} = r_{Cijm} + Z_{ij} \, \tau_m$, with the expectations, variances, covariances, and correlations among the Y_{im}'s given by $E\,(Y_{im}) = \tau_m$, $\text{var}\,(Y_{im}) = \omega_m^2$, $\text{cov}\,(Y_{im}, Y_{im'}) = \omega_{mm'}$ and $\text{corr}\,(Y_{im}, Y_{im'}) = \rho_{mm'} = \omega_{mm'} / (\omega_m \, \omega_{m'})$, so $\omega_m^2 = \omega_{mm}$. The covariance matrix $\boldsymbol{\Omega}$ is an $M \times M$ array with $\omega_{mm'}$ in row m and column m'. In this case, $Y_i = \sum_{m=1}^{M} \lambda_m \, Y_{im}$ is Normally distributed with expectation $\mu_{\boldsymbol{\lambda}} = \sum_{m=1}^{M} \lambda_m \, \tau_m$ and variance $\omega_{\boldsymbol{\lambda}}^2 = \sum_{m=1}^{M} \sum_{m'=1}^{M} \lambda_m \, \lambda_{m'} \omega_{mm'}$. Wilcoxon's signed rank statistic is calculated from the Y_i.

[2] A good reference for basic facts about the multivariate Normal distribution and associated matrix manipulations is C.R. Rao's text [9]. See, in particular, his definition of the multivariate Normal distribution [9, Chapter 8, Definition 1].

The situation is now the same as in Chap. 15, and the design sensitivity is

$$\widetilde{\Gamma} = \frac{p_1^{'}}{1 - p_1^{'}}, \tag{18.4}$$

where, for $i \neq i'$,

$$p_1^{'} = \Pr\left(Y_i + Y_{i'} > 0\right) = \Pr\left(\frac{Y_i + Y_{i'} - 2\mu_{\boldsymbol{\lambda}}}{\sqrt{2\omega_{\boldsymbol{\lambda}}^2}} > \frac{-2\mu_{\boldsymbol{\lambda}}}{\sqrt{2\omega_{\boldsymbol{\lambda}}^2}}\right) = \Phi\left(\frac{\sqrt{2}\mu_{\boldsymbol{\lambda}}}{\omega_{\boldsymbol{\lambda}}}\right). \tag{18.5}$$

Moreover, the optimal value, say $\boldsymbol{\lambda}_{\text{opt}}$, of $\boldsymbol{\lambda}$ with the largest design sensitivity $\widetilde{\Gamma}_{\text{opt}}$ is given by a simple formula involving the inverse of the covariance matrix, say $\boldsymbol{\Omega}$; specifically, $\boldsymbol{\lambda}_{\text{opt}}$ is any vector proportional to $\boldsymbol{\Omega}^{-1}\boldsymbol{\tau}$ where $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_M)^T$; see [4].[3]

18.2.3 Numerical Results for Bivariate Normal Responses

Table 18.1 computes the design sensitivities for bivariate Normal responses ($M = 2$), with an additive treatment effect (τ_1, τ_2), standard deviations (ω_1, ω_2), and correlation ρ_{12}; see also [4, 15] where $M > 2$ and other cases are considered. The table calculates four design sensitivities: (i) $\widetilde{\Gamma}_1$ and $\widetilde{\Gamma}_2$ are the design sensitivities for the two outcomes Y_{i1} and Y_{i2} used alone; (ii) $\widetilde{\Gamma}_=$ is the design sensitivity with equal weights, $\boldsymbol{\lambda} = (1, 1)$, analogous to the coherent signed rank statistic used in Table 5.3 ; (iii) $\widetilde{\Gamma}_{\text{opt}}$ is the optimal design sensitivity and $\boldsymbol{\lambda}_{\text{opt}}$ are the weights that produce the optimal design sensitivity. It should be kept in mind that the investigator does not know $\boldsymbol{\lambda}_{\text{opt}}$, and so the investigator can hope to get closer to $\widetilde{\Gamma}_{\text{opt}}$ in one way or another, but is unlikely to reach $\widetilde{\Gamma}_{\text{opt}}$.

In cases A–C, the effects (τ_1, τ_2) and standard deviations (ω_1, ω_2) are equal for the two outcomes and the optimal weights are equal, so $\widetilde{\Gamma}_1 = \widetilde{\Gamma}_2 \leq \widetilde{\Gamma}_{opt} = \widetilde{\Gamma}_=$. The advantage of using both outcomes with equal weights is substantial for uncorrelated outcomes, $\rho_{12} = 0$, and the advantage disappears as $\rho_{12} \to 1$.

Cases D–F resemble the situation in Chap. 16 in that the effects are equal, $\tau_1 = \tau_2$, but the standard deviation of Y_{i2} is half that of Y_{i1}; here, however, the investigator does not have to choose one outcome with a change in sample size, but can use both outcomes with the full sample size. In these cases, $\widetilde{\Gamma}_1 \leq \widetilde{\Gamma}_= \leq \widetilde{\Gamma}_2 \leq \widetilde{\Gamma}_{\text{opt}}$ and

[3]The proportionality constant cancels in (18.5). If $\boldsymbol{\lambda}_{\text{opt}} = \boldsymbol{\Omega}^{-1}\boldsymbol{\tau}$ is replaced by $2\boldsymbol{\lambda}_{\text{opt}}$, then both $\mu_{\boldsymbol{\lambda}}$ and $\omega_{\boldsymbol{\lambda}}$ are multiplied by 2, and (18.5) is unchanged. For this reason, it is often attractive to multiply $\boldsymbol{\lambda}_{\text{opt}}$ by a constant so that the values are easy to consider. Where possible, $\boldsymbol{\lambda}_{\text{opt}}$ is given as a vector of integers.

$\widetilde{\Gamma}_2 = \widetilde{\Gamma}_{opt}$ for $\rho_{12} = \frac{1}{2}$. The optimal procedure does not use the first outcome when $\rho_{12} = \frac{1}{2}$.

Cases G–I have equal standard deviations but unequal effects, $\tau_1 = \frac{3}{4}$, $\tau_2 = \frac{1}{4}$, so $\tau_1 + \tau_2$ is the same as in cases A–C, but τ_1 is larger. In these cases, $\widetilde{\Gamma}_2 \leq \widetilde{\Gamma}_= \leq \widetilde{\Gamma}_1 \leq \widetilde{\Gamma}_{\text{opt}}$, and the optimal procedure attaches negative weight to Y_{i2} for $\rho_{12} = \frac{1}{2}$.

In cases J–M, the second outcome is unaffected by treatment, with $\tau_2 = 0$. In these cases, using equal weights has inferior design sensitivity when compared to using the first outcome alone. If $\tau_2 = 0$ but $\rho_{12} > 0$, then the optimal weights, $\boldsymbol{\lambda}_{\text{opt}}$, attach a negative weight to the unaffected second outcome, and for $\rho_{12} = 0.75$ the increase in design sensitivity is large, $\widetilde{\Gamma}_{\text{opt}} = 6.0$ instead of $\widetilde{\Gamma}_1 = 3.2$. If $\tau_2 = 0$ but $\rho_{12} > 0$, then outcome two, Y_{i2}, is an unaffected outcome, or equivalently a control outcome, that is correlated with the first outcome, Y_{i1}, as discussed in Sect. 5.2.4 and [10, 11, 26]. So, cases K–M suggest one way that an unaffected outcome may be used so that it increases the design sensitivity. The simplest example of an unaffected Y_{i2} correlated with an affected Y_{i1} is a baseline, pretreatment measure of the same outcome. Notably, the optimal negative weight attached to the unaffected Y_{i2} is smaller in absolute value than the positive weight attached to Y_{i1}, so the optimal combination is not the difference-in-differences, $Y_{i1} - Y_{i2}$.

By definition, $\widetilde{\Gamma}_{\text{opt}}$ is always the largest design sensitivity. In all the cases considered in A–M, equal weights were better than the inferior of the two outcomes, $\min\left(\widetilde{\Gamma}_1, \widetilde{\Gamma}_2\right) \leq \widetilde{\Gamma}_=$. Equal weights, $\widetilde{\Gamma}_=$, were sometimes better and sometimes worse than the better of the two outcomes, that is, $\max\left(\widetilde{\Gamma}_1, \widetilde{\Gamma}_2\right)$. The optimal weights, $\widetilde{\Gamma}_{\text{opt}}$, were sometimes much better than the second best alternative, $\max\left(\widetilde{\Gamma}_1, \widetilde{\Gamma}_2, \widetilde{\Gamma}_=\right)$, and sometimes no better. The design sensitivities vary considerably in cases A–M, from 1.0 to 16.6, so knowing which circumstance will arise and how to adapt the analysis to that circumstance will materially affect the design sensitivity.

Cases A–M in Table 18.1 have various practical implications. In favorable circumstances, a coherent statistic with equal weights can yield materially less sensitivity to unmeasured covariates than an analysis that focuses on a single outcome; however, not all circumstances are favorable in this sense. Similar results in [4] show that, in favorable circumstances, the gains in $\widetilde{\Gamma}$ from using an equally weighted coherent statistic can increase substantially with additional outcomes, $M > 2$, and that the sturdier, more practical coherent signed rank statistic (Sect. 5.2.3, [12] and [13, §9]) exhibits performance similar to the simpler but less practical statistic evaluated in Table 18.1. At the same time, in Table 18.1, the optimal weights diminish the importance of an outcome that is either unstable (i.e., larger ω) or only slightly affected (i.e., smaller τ), perhaps ignoring such an outcome with zero weight, or perhaps even attaching negative weight to such an outcome; see Chap. 16 for a different but related phenomenon. Finally, the substantial variability in design sensitivities in cases A–M, from 1.0 to 16.6, suggests that a correct plan for analysis may materially affect the sensitivity of results to unmeasured biases, so time and resources should be allocated to planning the analysis; see Sect. 22.1.

Table 18.1 Design sensitivity with bivariate Normal outcomes and an additive treatment effect

Case	Standard deviation (ω_1, ω_2)	Correlation ρ_{12}	Effects (τ_1, τ_2)	Optimal weights $\boldsymbol{\lambda}_{\text{opt}}$	Optimal $\widetilde{\Gamma}_{\text{opt}}$	Outcome 1 $\widetilde{\Gamma}_1$	Outcome 2 $\widetilde{\Gamma}_2$	Equal weights $\widetilde{\Gamma}_=$
A	(1, 1)	0	(0.5, 0.5)	(1, 1)	5.3	3.2	3.2	5.3
B	(1, 1)	0.25	(0.5, 0.5)	(1, 1)	4.4	3.2	3.2	4.4
C	(1, 1)	0.5	(0.5, 0.5)	(1, 1)	3.8	3.2	3.2	3.8
D	(1, 0.5)	0	(0.5, 0.5)	(1, 4)	16.6	3.2	11.7	8.7
E	(1, 0.5)	0.25	(0.5, 0.5)	(1, 7)	12.9	3.2	11.7	7.1
F	(1, 0.5)	0.5	(0.5, 0.5)	(0, 1)	11.7	3.2	11.7	6.0
G	(1, 1)	0	(0.75, 0.25)	(3, 1)	6.6	5.9	1.8	5.3
H	(1, 1)	0.25	(0.75, 0.25)	(11, 1)	6.0	5.9	1.8	4.4
I	(1, 1)	0.5	(0.75, 0.25)	(5, −1)	6.1	5.9	1.8	3.8
J	(1, 1)	0.0	(0.5, 0.0)	(1, 0)	3.2	3.2	1.0	2.2
K	(1, 1)	0.25	(0.5, 0.0)	(4, −1)	3.3	3.2	1.0	2.1
L	(1, 1)	0.5	(0.5, 0.0)	(2, −1)	3.8	3.2	1.0	1.9
M	(1, 1)	0.75	(0.5, 0.0)	(4, −3)	6.0	3.2	1.0	1.8

Four design sensitivities are given: for optimal weights, for outcome 1 alone, for outcome 2 alone, and for equal weights. The optimal weights are also given

18.2.4 Practical Implementation for General $\boldsymbol{\lambda}$

The statistic considered in Sect. 18.2 is easy to study theoretically: it applies Wilcoxon's signed rank statistic to a linear combination $\boldsymbol{\lambda}$ of outcomes. For instance, there is a simple formula (18.4) for the design sensitivity $\widetilde{\Gamma}$, leading to the calculations in Table 18.1. The statistic is less useful in practice, because the interpretation of $\boldsymbol{\lambda}$ depends upon the scale of measurement for the M outcomes. For instance, $\boldsymbol{\lambda} = (1, 1)$ can be understood as giving equal weight to $M = 2$ outcomes only if the two outcomes have the same "scale" of measurement, in some suitable sense.

Instead of applying Wilcoxon's statistic to a linear combination of outcomes, the coherent signed rank statistic in Sect. 5.2.3 is the sum of M signed rank statistics. The absolute ranks used in Wilcoxon's statistic placed the several outcomes on a common scale, and did this in a manner unaffected by outliers. The practical implementation for general $\boldsymbol{\lambda}$ takes as its test statistic a weighted combination of M signed rank statistics, with weight λ_m for the statistic computed from outcome m in (18.3), $m = 1, \ldots, M$.

Let q_{im} be the rank of the absolute pair difference $|Y_{im}|$ for outcome m in (18.3), setting $q_{im} = 0$ if $|Y_{im}| = 0$. Define $\text{sign}(y)$ as 1 if $y > 0$, 0 if $y = 0$, or -1 if $y < 0$. Notice that $\text{sign}(y)$ differs slightly from $\text{sgn}(y)$ in Sect. 2.3.3, where $\text{sgn}(y)$ appears in the usual definition of Wilcoxon's signed rank statistic. The centered form of Wilcoxon's signed rank statistic [7] for outcome m is $T_m = \sum_{i=1}^{I} \text{sign}(Y_{im}) q_{im}$. This centering does not change the consequential behavior of the signed rank statistic, but it leads to tidier notation in the following discussion.

For $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_M)^T \neq \mathbf{0}$, define a weighted combination of centered signed rank statistics, $T_{\boldsymbol{\lambda}} = \sum_{m=1}^{M} \lambda_m T_m$. Essentially, the coherent signed rank statistic in Sect. 5.2.3 is $T_{\boldsymbol{\lambda}}$ with $\boldsymbol{\lambda} = (1, 1, \ldots, 1)^T$. Conveniently, $T_{\boldsymbol{\lambda}}$ is not a radically new statistic, but merely another signed rank statistic with different rank scores, as will now be demonstrated.

The observed value R_{ijm} of outcome m is r_{Tijm} if individual ij receives treatment with $Z_{ij} = 1$, or it is r_{Cijm} if individual ij receives control, with $Z_{ij} = 0$. Under H_0, these two potential values of each outcome are equal, $r_{Cijm} = r_{Tijm}$ for all i, j, and m. With M outcomes, $\mathscr{F}$ includes the potential outcomes, r_{Tijm} and r_{Cijm} for all m.

Define $c_{im} = \text{sign}(R_{i1m} - R_{i2m})$, noting that $c_{im} = \text{sign}(r_{Ci1m} - r_{Ci2m})$ if H_0 is true. Using (18.3), if H_0 is true, then $Y_{im} = (Z_{i1} - Z_{i2})\,(r_{Ci1m} - r_{Ci2m})$, so that $\text{sign}\,(Y_{im}) = \text{sign}\,(Z_{i1} - Z_{i2})\, c_{im}$ and

$$\text{sign}\,(Y_{im})\; q_{im} = \text{sign}\,(Z_{i1} - Z_{i2})\; c_{im}\, q_{im}. \tag{18.6}$$

Write $q_{\boldsymbol{\lambda} i} = \sum_{m=1}^{M} \lambda_m \, c_{im} \, q_{im}$. Using (18.6), we may rewrite $T_{\boldsymbol{\lambda}}$ as another centered signed rank statistic, namely

$$T_{\boldsymbol{\lambda}} = \sum_{m=1}^{M} \lambda_m \, T_m = \sum_{i=1}^{I} \sum_{m=1}^{M} \lambda_m \, \mathrm{sign}(Y_{im}) \, q_{im} = \sum_{i=1}^{I} \mathrm{sign}\,(Z_{i1} - Z_{i2}) \; q_{\boldsymbol{\lambda} i}. \tag{18.7}$$

Because $T_{\boldsymbol{\lambda}}$ is just a centered version of another signed rank statistic, small adjustments to the method in Sect. 3.5 yield its sensitivity analysis.

The formulas for the sensitivity analysis for $T_{\boldsymbol{\lambda}}$ differ slightly from those in Sect. 3.5 for two reasons. First, the $q_{\boldsymbol{\lambda} i}$ are not a rearrangement of the numbers $1, 2, \ldots, I$, so the formulas for $q_{\boldsymbol{\lambda} i}$ resemble the untidy formulas for ties in Sect. 3.5, not the tidy formulas for untied responses. Second, centering implies $\mathrm{sign}\,(Z_{i1} - Z_{i2}) \, q_{\boldsymbol{\lambda} i}$ takes the values $\pm q_{\boldsymbol{\lambda} i}$, not 0 or $|q_{\boldsymbol{\lambda} i}|$.

Define $\overline{\overline{T}}_{\boldsymbol{\lambda}}$ to be the sum of I independent random variables, where the ith random variable is $|q_{\boldsymbol{\lambda} i}|$ with probability $\Gamma/(1+\Gamma)$ or $-|q_{\boldsymbol{\lambda} i}|$ with probability $1/(1+\Gamma)$ if $|q_{\boldsymbol{\lambda} i}| \neq 0$, and is 0 with probability one if $|q_{\boldsymbol{\lambda} i}| = 0$. If H_0 is true and the bias in treatment assignment is at most Γ, then $\overline{\overline{T}}_{\boldsymbol{\lambda}}$ is too large: it is stochastically larger than $T_{\boldsymbol{\lambda}}$.

When H_0 is true, the expectation of $\overline{\overline{T}}_{\boldsymbol{\lambda}}$ is

$$E\left(\overline{\overline{T}}_{\boldsymbol{\lambda}} \,\middle|\, \mathscr{F}, \mathscr{Z}\right) = \frac{\Gamma - 1}{1 + \Gamma} \sum_{i=1}^{I} |q_{\boldsymbol{\lambda} i}|. \tag{18.8}$$

In parallel, when H_0 is true, the variance of $\overline{\overline{T}}_{\boldsymbol{\lambda}}$ is

$$\mathrm{var}\left(\overline{\overline{T}}_{\boldsymbol{\lambda}} \,\middle|\, \mathscr{F}, \mathscr{Z}\right) = \frac{4\Gamma}{(1+\Gamma)^2} \sum_{i=1}^{I} q_{\boldsymbol{\lambda} i}^2. \tag{18.9}$$

The deviate

$$Q_{\boldsymbol{\lambda}} = \frac{T_{\boldsymbol{\lambda}} - \frac{\Gamma-1}{1+\Gamma} \sum_{i=1}^{I} |q_{\boldsymbol{\lambda} i}|}{\sqrt{\frac{4\Gamma}{(1+\Gamma)^2} \sum_{i=1}^{I} q_{\boldsymbol{\lambda} i}^2}} \tag{18.10}$$

is computed, and $1 - \Phi\,(Q_{\boldsymbol{\lambda}})$ is the upper bound on the approximate one-sided P-value testing H_0 in the presence of a bias in treatment assignment of at most Γ.

Consider, again, the coherence of the results of the study by Angrist and Lavy [1] as discussed in Sect. 5.2.3. The deviate for math scores alone is (18.10) with $\boldsymbol{\lambda} = (1, 0)^T$. The deviate for verbal scores alone is (18.10) with $\boldsymbol{\lambda} = (0, 1)^T$. The deviate for coherent signed rank statistic is (18.10) with $\boldsymbol{\lambda} = (1, 1)^T$. At $\Gamma = 1.65$ in Table 5.2, the deviate is $Q_{\boldsymbol{\lambda}} = 1.0913$ with P-value bound $1 - \Phi(1.0913) = 0.138$

for math alone, $\boldsymbol{\lambda} = (1, 0)^T$, is $Q_{\boldsymbol{\lambda}} = 1.442$ with P-value bound $1 - \Phi(1.442) = 0.075$ for verbal alone, $\boldsymbol{\lambda} = (0, 1)^T$, but is $Q_{\boldsymbol{\lambda}} = 1.713$ with P-value bound $1 - \Phi(1.713) = 0.043$ for the coherent statistic with $\boldsymbol{\lambda} = (1, 1)^T$.

18.3 Can Coherence Be Discovered?

Can you discover a coherent pattern of responses in the data at hand, or must you anticipate the pattern? If you knew $\boldsymbol{\lambda}_{\text{opt}}$ in Table 18.1, then in each situation the design sensitivity would be $\widetilde{\Gamma}_{\text{opt}}$, which is often larger, sometimes much larger, than the design sensitivity for a single outcome. How could you know, without looking at the data, that $\boldsymbol{\lambda}_{\text{opt}}$ would be $(2, -1)^T$, as it is in case L of Table 18.1?

18.3.1 Using Split Samples to Plan for Coherence

A simple strategy randomly divides the I matched pairs into two groups of pairs, planning the analysis using one group and implementing it on the other; see Sect. 22.1 and [3]. The weights $\boldsymbol{\lambda}$ are suggested by one group of pairs and then applied as a fixed weights in the other. Heller et al. [4] consider a split such as $I/10$ pairs to plan the analysis, to find a good $\boldsymbol{\lambda}$, followed by an analysis of the remaining $9I/10$ pairs with this fixed $\boldsymbol{\lambda}$. Zhang et al. [30] use split samples to improve the design of an observational study with about $I = 130{,}000$ pairs, so $I/10 = 13{,}000$. See also Zhao et al. [31] for a different approach to sample splitting.

The design sensitivity, $\widetilde{\Gamma}$, is a limit as $I \to \infty$, so it is unaffected by the reduced sample size in splitting strategies. In practice, however, the reduced sample size is a consideration. As $I \to \infty$, a planning sample of $I/10$ pairs suffices to develop a good plan, and a complementary sample of $9I/10$ pairs suffices for analysis; however, for finite I, either sample, or both samples, may be too small to accomplish their respective tasks. Split samples can be attractive when many planning decisions are needed, including subjective decisions that are difficult to study formally. In contrast, if the only task is to choose $\boldsymbol{\lambda}$, then we can avoid sample splitting and use all of the data with corrections for multiple testing that will now be described.

18.3.2 Considering Every Possible $\boldsymbol{\lambda}$

Rather than split the sample to improve our guess about a good choice of $\boldsymbol{\lambda}$, a general technique due to Scheffé [23] permits consideration of every possible $\boldsymbol{\lambda}$ with a correction for testing infinitely many $\boldsymbol{\lambda}$'s. In effect, the test statistic is the maximum value of $Q_{\boldsymbol{\lambda}}$ in (18.10), or $\max_{\boldsymbol{\lambda} \neq \mathbf{0}} Q_{\boldsymbol{\lambda}}$. The resulting procedure pays a

price for multiple testing, but it achieves the optimal design sensitivity, $\widetilde{\Gamma}_{\text{opt}}$, that is, the largest design sensitivity that can be produced by a choice of $\boldsymbol{\lambda}$.

What does the correction for multiple testing achieve? We wish to control the probability of rejecting a true null hypothesis of no treatment effect at level α, conventionally $\alpha = 0.05$. With a single outcome, a sensitivity analysis performed at a given value of Γ falsely rejects no effect with probability at most α providing the bias in treatment assignment is at most Γ. With M outcomes instead of one, the one null hypothesis, H_0, asserts that the treatment has no effect on these outcomes, but that one hypothesis is tested in infinitely many ways, one for each $\boldsymbol{\lambda} \neq \mathbf{0}$. The correction for multiple testing makes a promise about the probability that at least one $\boldsymbol{\lambda} \neq \mathbf{0}$ leads to a false rejection of the hypothesis H_0 of no effect. If H_0 is true, the probability is at least $1 - \alpha$ that all $\boldsymbol{\lambda} \neq \mathbf{0}$ fail to reject H_0 when the bias in treatment assignment is at most Γ. The investigator can pick one $\boldsymbol{\lambda}$ after looking at the data, or can try every possible $\boldsymbol{\lambda} \neq \mathbf{0}$, yet be assured that the probability of a false rejection is at most α.

How large is the correction for multiple testing? Consider the large sample case, as $I \to \infty$, so that we can approximate the distribution of Wilcoxon's statistic by a Normal distribution as a consequence of the central limit theorem. For a single outcome or for a single $\boldsymbol{\lambda}$ chosen in advance, in a one-sided test, the deviate in (3.23) leads to rejection at level $\alpha = 0.05$ if it exceeds the critical value 1.65. In a two-sided test, the critical value is 1.96. Because the multiple testing procedure considers all possible $\boldsymbol{\lambda}$, it considers both $\boldsymbol{\lambda}$ and $-\boldsymbol{\lambda}$; so, in that limited sense, it is two-sided, or perhaps all-sided. At $\alpha = 0.05$, for $M = 2$ outcomes, the Scheffé method will reject if there exists even one $\boldsymbol{\lambda}$ such that the deviate analogous to (3.23) exceeds 2.45, rather than 1.65 or 1.96. With $M = 3$ outcomes, the deviate must exceed 2.80 for some $\boldsymbol{\lambda}$. In general, to reject H_0, there must be at least one $\boldsymbol{\lambda}$ whose deviate equals or exceeds the square root of the upper 1-α percentage point of the chi-square distribution on M degrees of freedom. So, a nontrivial but limited price is paid for an unlimited exploration of all $\boldsymbol{\lambda} \neq \mathbf{0}$.

Stated a little more precisely, let $\chi^2_{\alpha,M}$ be the upper α percentage point of the chi-square distribution on M degrees of freedom, where there are M outcomes. Under H_0, if the bias in treatment assignment is at most Γ, consider the probability $\varpi_I = \Pr\left(\max_{\boldsymbol{\lambda}\neq\mathbf{0}} Q_{\boldsymbol{\lambda}} \geq \sqrt{\chi^2_{\alpha,M}}\right)$ in a sample of I pairs. If $\epsilon > 0$, then $\varpi_I \leq \alpha + \epsilon$ for sufficiently large I; see [20, Proposition 1].

The method is easily seen to achieve the maximum design sensitivity, $\widetilde{\Gamma}_{\text{opt}}$; see [20, §4.3]. We do not know $\boldsymbol{\lambda}_{\text{opt}}$, but if we try every $\boldsymbol{\lambda}$ we will eventually try $\boldsymbol{\lambda}_{\text{opt}}$. If the sensitivity analysis is done with $\Gamma < \widetilde{\Gamma}_{\text{opt}}$, then the deviate at $\boldsymbol{\lambda}_{\text{opt}}$ tends to ∞ as $I \to \infty$, so eventually the deviate exceeds the critical constant that has been adjusted for multiple testing, say 2.45 for $M = 2$ and $\alpha = 0.05$. That is, the power of Scheffé's method tends to one as $I \to \infty$ for all $\Gamma < \widetilde{\Gamma}_{\text{opt}}$.

For fixed $\Gamma > 1$, consider the power of a sensitivity analysis based on $\max_{\boldsymbol{\lambda}\neq\mathbf{0}} Q_{\boldsymbol{\lambda}}$ in favorable situations such as those in Table 18.1. Because $\max_{\boldsymbol{\lambda}\neq\mathbf{0}} Q_{\boldsymbol{\lambda}}$ achieves the maximum design sensitivity, $\widetilde{\Gamma}_{\text{opt}}$, we expect it to perform well for large I as Γ is increased towards $\widetilde{\Gamma}_{\text{opt}}$. That is, for a fixed $\boldsymbol{\lambda}$, the fixed statistic $Q_{\boldsymbol{\lambda}}$

has design sensitivity $\widetilde{\Gamma}_{\boldsymbol{\lambda}}$. If $\widetilde{\Gamma}_{\boldsymbol{\lambda}} < \widetilde{\Gamma}_{\text{opt}}$, then for Γ strictly between $\widetilde{\Gamma}_{\boldsymbol{\lambda}}$ and $\widetilde{\Gamma}_{\text{opt}}$, as $I \to \infty$, the power of $Q_{\boldsymbol{\lambda}}$ is tending to zero while the power of $\max_{\boldsymbol{\lambda} \neq \mathbf{0}} Q_{\boldsymbol{\lambda}}$ is tending to one. For a simulation of power in finite samples, see [20, Table 3].

As the number M of outcomes increases, Sheffé's method pays an ever larger price for considering every possible $\boldsymbol{\lambda}$. Therefore, the method is useful for small M, say $M = 2$ or $M = 3$, and fairly large I. We seem to face the uncomfortable choice of guessing $\boldsymbol{\lambda}$, perhaps guessing incorrectly, or paying to consider every $\boldsymbol{\lambda}$ when not all $\boldsymbol{\lambda}$'s seem equally reasonable or interesting. Is there a third option?

18.3.3 Hedged Bets About $\boldsymbol{\lambda}$

An alternative [21] to Scheffé's method combines one fixed comparison, say $\boldsymbol{\kappa}$, chosen a priori, with exhaustive consideration of all possible $\boldsymbol{\lambda}$. In effect, we place a hedged bet: we bet something on our guess, $\boldsymbol{\kappa}$, but we reserve the right to consider every $\boldsymbol{\lambda}$. In the simplest case, $\boldsymbol{\kappa} = (1, 1, \ldots, 1)^T$ yielding the coherent signed rank statistic in Sect. 5.2.3, so we bet that all M outcomes are similarly affected by treatment; however, we retain the right to look at the M outcomes one at time, as well as the $2^M - 1$ subsets of the outcomes, and other comparisons.

The method works with the joint distribution of $Q_{\boldsymbol{\kappa}}$ and $\max_{\boldsymbol{\lambda} \neq \mathbf{0}} Q_{\boldsymbol{\lambda}}$, but gives equal emphasis to these two statistics. As a result, $Q_{\boldsymbol{\kappa}}$ is compared to a smaller critical value and $\max_{\boldsymbol{\lambda} \neq \mathbf{0}} Q_{\boldsymbol{\lambda}}$ is compared to a larger critical value.

Consider the case of $M = 2$ outcomes in a level $\alpha = 0.05$ test. If we focused on $Q_{\boldsymbol{\kappa}}$ alone having chosen $\boldsymbol{\kappa}$ a priori, then we reject H_0 if $Q_{\boldsymbol{\kappa}} \geq 1.65$ for a one-tailed test, or if $|Q_{\boldsymbol{\kappa}}| \geq 1.96$ for a two-tailed test. If we consider $Q_{\boldsymbol{\kappa}}$ and $\max_{\boldsymbol{\lambda} \neq \mathbf{0}} Q_{\boldsymbol{\lambda}}$ jointly, then we reject H_0 if $Q_{\boldsymbol{\kappa}} \geq 1.89$ or if $\max_{\boldsymbol{\lambda} \neq \mathbf{0}} Q_{\boldsymbol{\lambda}} \geq 2.66$. Implicity, we have done a two-sided test for $\boldsymbol{\kappa}$, because $\max_{\boldsymbol{\lambda} \neq \mathbf{0}} Q_{\boldsymbol{\lambda}}$ does consider $-\boldsymbol{\kappa}$ as one of infinitely many $\boldsymbol{\lambda}$'s, but we have emphasized one tail of $Q_{\boldsymbol{\kappa}}$ at the expense of the other. Each of the two statistics, $Q_{\boldsymbol{\kappa}}$ and $\max_{\boldsymbol{\lambda} \neq \mathbf{0}} Q_{\boldsymbol{\lambda}}$, has a probability of 0.029 of exceeding its critical value, but the chance that at least one statistic exceeds its critical value is $0.050 < 0.029 + 0.029 = 0.058$. We could have used the Bonferroni inequality to correct for using both $Q_{\boldsymbol{\kappa}}$ and $\max_{\boldsymbol{\lambda} \neq \mathbf{0}} Q_{\boldsymbol{\lambda}}$, but that approach would have been conservative, with larger critical constants for both statistics. Scheffé's method does not distinguish $Q_{\boldsymbol{\kappa}}$ and compares every $Q_{\boldsymbol{\lambda}}$ to the same critical constant, namely 2.45, which is slightly below 2.66 but well above 1.89.

Like Scheffé's method, as the number of outcomes M increases, the critical constant for $\max_{\boldsymbol{\lambda} \neq \mathbf{0}} Q_{\boldsymbol{\lambda}}$ increases. However, as M increases, the critical constant for $Q_{\boldsymbol{\kappa}}$ barely changes, staying below the upper critical value, $\Phi^{-1}(1 - \alpha/2)$, for a two-sided α-level test, that is, below $\Phi^{-1}(1 - 0.025) = 1.96$ for $\alpha = 0.05$. Stated informally, the cost of betting on $\boldsymbol{\kappa}$ barely increases with M, but the cost of hedging that bet does increase with M.

Like Scheffé's method, the hedged bet also achieves the largest design sensitivity, $\widetilde{\Gamma}_{\text{opt}} = \max_{\boldsymbol{\lambda} \neq \mathbf{0}} \widetilde{\Gamma}_{\boldsymbol{\lambda}}$ by essentially the same argument [21, §6].

Consider the situation with $M = 2$ in Table 5.2 with $\Gamma = 1.5$. The coherent signed rank statistic, $Q_{\boldsymbol{\kappa}}$ with $\boldsymbol{\kappa} = (1, 1)^T$, has $Q_{\boldsymbol{\kappa}} = 2.054$, so the hedged bet leads to rejection of H_0 at the 0.05 level, because $2.054 > 1.89$, but other comparisons $Q_{\boldsymbol{\lambda}}$ do not reject when compared to 2.66. Scheffé's method does not reject for any $Q_{\boldsymbol{\lambda}}$, because it compares $Q_{\boldsymbol{\kappa}}$ to 2.45, not to 1.89.

The guess, $\boldsymbol{\kappa}$, must be chosen a priori, that is, without using Z_{ij}; however, $\boldsymbol{\kappa}$ may depend on quantities that are fixed when H_0 is true [21, §4.3]. For instance, if H_0 were true, then $R_{ijm} = r_{Cijm}$ is fixed by conditioning on $\mathscr{F}$ in permutation tests, so $\boldsymbol{\kappa}$ could be set to the weights that define the first principal component of the R_{ijm}. Indeed, $\max_{\boldsymbol{\lambda} \neq \mathbf{0}} Q_{\boldsymbol{\lambda}}$ could be restricted to the space spanned by the first, say three principal components, so there is a hedged bet that the treatment affects the first principal component, with an exhaustive search for effects in the space spanned by the first three components. Because $\boldsymbol{\lambda}$ is restricted to the first three principal components, critical constants are smaller: they are from the chi-square distribution on 3 degrees of freedom, rather than M degrees of freedom, a substantial difference if the number of outcomes M is moderately large. Alternatively, one might use principal components of $c_{im}\, q_{im}$ in Sect. 18.6. See [21] for an example involving three measures of cognitive decline in the elderly.

18.3.4 Summary

Coherence among M outcomes can materially increase the design sensitivity, $\widetilde{\Gamma}$, providing $\boldsymbol{\lambda}$ is wisely chosen. In ignorance of the best $\boldsymbol{\lambda}$, we may still increase design sensitivity by testing every $\boldsymbol{\lambda}$ while correcting for testing many hypotheses. Perhaps best of all, we may place a hedged bet on one comparison, $\boldsymbol{\kappa}$, while also considering every $\boldsymbol{\lambda}$, paying only a modest price for hedging the bet. When the number, M, of outcomes, R_{ijm}, is not very small, we may bet on low dimensional summaries of those outcomes, such as the first few principal components defined without reference to the treatment assignments Z_{ij}.

18.4 Doses

Recall from Sect. 5.2.5 that there are conflicting claims in the literature about whether or not dose–response relationships are important in appraising evidence about effects caused by a treatment. Also, recall the example in Table 5.5 in which doses did seem to reduce sensitivity to unmeasured biases. Calculations of design sensitivity in the current section will shed light on the various conflicting claims. In these calculations, the Y_i will be drawn from a continuous distribution, so ties do not occur.

18.4.1 Another Way to Write a Signed Rank Statistic

Wilcoxon's signed rank statistic is $T = \sum_{i=1}^{I} \operatorname{sgn}(Y_i)\, q_i$, where q_i is the rank of $|Y_i|$, and the dose-weighted signed rank statistic [12, 14, 27] is $T_{\text{dose}} = \sum_{i=1}^{I} \operatorname{sgn}(Y_i)\, q_i\, d_i$, where $d_i > 0$ is the dose of treatment applied to the treated subject in pair i.[4] In calculating the design sensitivity, $\widetilde{\Gamma}$, it is helpful to express T and T_{dose} in an equivalent but slightly different form. The reason for this is that the expectation of T_{dose} will be needed in the "favorable situation," and the expectation has a simple form when T_{dose} is written in a different form.

Write

$$W_{ik} = \begin{cases} 1 \text{ if } \ |Y_i| \geq |Y_k| \ \text{ and } \ Y_i > 0 \\ 0 \text{ otherwise.} \end{cases} \tag{18.11}$$

It takes only a moment to realize that the dose-weighted signed rank statistic, $T_{\text{dose}} = \sum_{i=1}^{I} \operatorname{sgn}(Y_i)\, q_i\, d_i$, is equal to $\sum_{i=1}^{I} \sum_{k=1}^{I} d_i W_{ik}$.[5] When all of the doses, d_i, equal 1, this simplifies to say that Wilcoxon's signed rank statistic $T = \sum_{i=1}^{I} \operatorname{sgn}(Y_i)\, q_i$ equals $\sum_{i=1}^{I} \sum_{k=1}^{I} W_{ik}$. The advantage of writing T_{dose} in this way is that it yields a simple formula for $E(T_{\text{dose}})$, namely $E(T_{\text{dose}}) = \sum_{i=1}^{I} \sum_{k=1}^{I} E(d_i W_{ik})$.

Another similar quantity, V_{ik}, is also needed. Write

$$V_{ik} = \begin{cases} 1 \text{ if } \ |Y_i| \geq |Y_k| \\ 0 \text{ otherwise.} \end{cases} \tag{18.12}$$

Unlike W_{ik} in (18.11), the quantity V_{ik} ignores the sign of Y_i. In takes only another moment to realize that $q_i = \sum_{k=1}^{I} V_{ik}$ and $d_i\, q_i = \sum_{k=1}^{I} d_i V_{ik}$.[6] Obviously, $W_{ik} \leq V_{ik}$.

[4]Recall footnote 8 of Sect. 5.2.5.

[5]To see this, compare $d_i \operatorname{sgn}(Y_i)\, q_i$ to $\sum_{k=1}^{I} d_i W_{ik}$. If $Y_i \leq 0$, then $\operatorname{sgn}(Y_i) = 0$ so $d_i \operatorname{sgn}(Y_i)\, q_i = 0$, but also $d_i W_{ik} = 0$ for every k, so $0 = \sum_{k=1}^{I} d_i W_{ik}$. If $Y_i > 0$, then $\operatorname{sgn}(Y_i) = 1$, so $d_i \operatorname{sgn}(Y_i)\, q_i = d_i\, q_i$ is d_i times the rank q_i of $|Y_i|$. Also $d_i \sum_{k=1}^{I} W_{ik}$ is d_i times the number of $|Y_k|$ such that $|Y_i| \geq |Y_k|$, which is, of course, just d_i times the rank q_i of $|Y_i|$. So for every i, $d_i \operatorname{sgn}(Y_i)\, q_i = \sum_{k=1}^{I} d_i W_{ik}$. This shows that $T_{\text{dose}} = \sum_{i=1}^{I} \sum_{k=1}^{I} d_i W_{ik}$. If you set $d_i = 1$ for all i, then this becomes $T = \sum_{i=1}^{I} \sum_{k=1}^{I} W_{ik}$.

[6]As in the previous footnote, $\sum_{k=1}^{I} V_{ik}$ is the number of $|Y_k|$ that are less than or equal to $|Y_i|$, which is just the rank q_i of $|Y_i|$.

18.4.2 The Favorable Situation with Doses

As in Chap. 15, we ask: Will we be in a position to report that the conclusions are insensitive to unmeasured bias when, indeed, the treatment did have an effect and there was no bias from unmeasured covariates? It is in this situation that we hope to report that the results are insensitive to small and moderate biases. In other words, it is assumed that subjects who look comparable in terms of observed covariates, $\mathbf{x}_{ij}$, are indeed comparable, so the naïve model (3.5)–(3.8)is correct. In addition, a model is needed in which there is a treatment effect. Two such models will be considered, one a special case of the other more general case.

In the general case of the "favorable situation," there are $L \geq 1$ possible positive doses, $\delta_1, \ldots, \delta_L$. Of course, $L = 1$ represents the situation "without doses" in the sense that every treated subject receives the same dose. The I doses d_i for the I matched pairs are independently sampled from the L possible doses, $\delta_1, \ldots, \delta_L$ with probabilities $\Pr(d_i = \delta_\ell) = \varphi_\ell$ for $\ell = 1, \ldots, L$; in effect, the doses are determined by sampling a multinomial distribution with L categories. Conditionally given that $d_i = \delta_\ell$, the treated-minus-control difference in observed responses is $Y_i = \tau_\ell + \epsilon_i$ where ϵ_i are independently sampled from a continuous distribution $F_\ell(\cdot)$ that is symmetric about zero. That is, for each fixed dose, the "favorable situation" is the same as in Chap. 15, namely an additive constant treatment effect τ_ℓ with independent, identically distributed symmetric, continuous errors, ϵ_i; in particular, with $L = 1$ dose, the situation is identical to that in Chap. 15. The new element is that the size of the treatment effect, τ_ℓ, may vary with the dose, $d_i = \delta_\ell$, and so may the error distribution, $F_\ell(\cdot)$. The model as just described first samples doses, $d_i = \delta_\ell$, then samples outcome differences, Y_i, conditionally given doses, but this implies a joint sampling scheme in which the dose outcome pairs (d_i, Y_i) are independent, identically distributed bivariate measurements (iid).

In this general model, two quantities, Ψ and Λ, will now be defined that are important for the design sensitivity, $\widetilde{\Gamma}$. Pick an i and a k with $i \neq k$. Because the bivariate (d_i, Y_i) are iid, it makes no difference which i and k are used, provided they are different. Define $\Psi = E(d_i W_{ik})$ and $\Lambda = E(d_i V_{ik})$, where W_{ik} is defined in (18.11) and V_{ik} is defined in (18.12). For specified $\delta_1, \ldots, \delta_L$, $\tau_1, \ldots, \tau_L$, and $F_1(\cdot), \ldots, F_L(.)$, the quantities Ψ and Λ are well defined; for instance, they could be determined by Monte Carlo.[7] Because $d_i W_{ik} \leq d_i V_{ik}$, it follows that $\Psi \leq \Lambda$; moreover, if $\Pr(Y_i > 0) < 1$, then $\Psi < \Lambda$.

In addition to this general model, it is helpful to have a more tangible, more specific model to permit tabulation of a few numerical results. In this model, the doses are $\delta_1 = 1, \delta_2 = 2, \ldots, \delta_L = L$, and they have equal probabilities, $\varphi_\ell = 1/L$ for $\ell = 1, \ldots, L$. Then the expected and median dose are both $(L+1)/2$. Also,

[7] You can write Ψ or Λ as a weighted sum over the L^2 possible dose combinations in pairs i and k, and in simple situations the individual terms in this sum either have closed forms or can be evaluated numerically. For some details, see [17]. These formulas are an aid to numerical computation, but they provide no insights, so they are not presented here.

the effect is linear in the dose, $\tau_\ell = \beta\delta_\ell$ and the error distribution $F_\ell(\cdot)$ is the same $F(\cdot)$ for all ℓ. The median treatment effect is then $\beta(L+1)/2$, and this is also the expected treatment effect if $F(\cdot)$ has finite expectation. If the median effect, $\beta(L+1)/2$, is held fixed, we may ask: Is it better to have more dose categories, L? If you could choose between a somewhat larger median effect, $\beta(L+1)/2$, with $L = 1$ dose category, or a somewhat smaller median effect, $\beta(L+1)/2$, with several dose categories (larger L), which would be the better choice? How, if at all, does the sample size, I, affect these choices? Before looking at numerical results, in speculating about these questions, one might look again at the example in Table 5.5.

18.4.3 A Formula for the Design Sensitivity with Doses

The following proposition gives the design sensitivity of the dose-weighted signed rank statistic T_{dose}. The proof is in the appendix to this chapter. Recall that $\Lambda \geq \Psi$ with equality if and only if $\Pr(Y_i > 0) = 1$.

Proposition 18.1 *Suppose that (i) doses d_i in the I pairs are independently sampled from a discrete distribution with $L \geq 1$ possible positive doses, $\delta_1, \ldots, \delta_L$ with $\Pr(d_i = \delta_\ell) = \varphi_\ell$, $\ell = 1, \ldots, L$; (ii) there is no bias from unmeasured covariates in the sense that conditions (3.5)–(3.8) are true, (iii) conditionally given that $d_i = \delta_\ell$, the treated-minus-control difference in observed responses is $Y_i = \tau_\ell + \epsilon_i$ where ϵ_i are independently sampled from a continuous distribution $F_\ell(\cdot)$ which is symmetric about zero. Then, as $I \to \infty$, the power of a sensitivity analysis using the dose-weighted signed rank statistic, T_{dose}, tends to 1 for $\Gamma < \widetilde{\Gamma}$ and to 0 for $\Gamma > \widetilde{\Gamma}$ where*

$$\widetilde{\Gamma} = \frac{\Psi}{\Lambda - \Psi}, \tag{18.13}$$

$\Psi = E(d_i W_{ik})$, and $\Lambda = E(d_i V_{ik})$ with $i \neq k$.

Proposition 18.1 generalizes Proposition 15.1 of Chap. 15 in the following sense. Consider the case of a single dose, $L = 1$, with $d_i = 1$ for all i. With such a single dose, the dose-weighted signed rank statistic T_{dose} equals Wilcoxon's signed rank statistic T. In this case, $\Lambda = 1/2$ and $\Psi = p_1^{'}/2 = \Pr(Y_i + Y_k > 0)/2$; so Proposition 18.1 implies Proposition 15.1 when there is a single dose.[8]

[8]To see this, assume that $d_i = 1$ for every i. Observe that with continuously distributed Y_i, there is zero probability that $|Y_i| = |Y_k|$ with $i \neq k$. Because Y_i and Y_k are two independent observations from the same continuous distribution, $\Pr(|Y_i| > |Y_k|) = \Pr(|Y_i| < |Y_k|) = 1/2$. Therefore, because $d_i = 1$, $\Lambda = E(d_i V_{ik}) = E(V_{ik}) = 1/2$. Also, $Y_i + Y_k > 0$ can occur in two mutually exclusive ways, either $|Y_i| > |Y_k|$ and $Y_i > 0$ or $|Y_k| > |Y_i|$ and $Y_k > 0$, and each of these

Recall that $\Lambda \geq \Psi$ with equality if and only if $\Pr(Y_i > 0) = 1$; in this case, correctly, $\widetilde{\Gamma} = \infty$ in (18.13).

18.4.4 Numerical Evaluation of the Design Sensitivity

Table 18.2 provides numerical values of the design sensitivity $\widetilde{\Gamma}$ from (18.13) in the case of L equally probable, equally spaced integer doses, $\delta_1 = 1, \ldots, \delta_L = L$, effect $\tau_\ell = \beta\,\delta_\ell$ at dose δ_ℓ with errors that are Normal, logistic, or Cauchy. The median treatment effect is $\beta\,(L+1)\,/2$.

In cases A through D in Table 18.2, the median treatment effect is $\beta\,(L+1)\,/2 = 1/2$, but the number of dose levels varies from 1 in case A to 9 in case D. That is, the dose and the effect are constant at $1/2$ in case A, but the doses range from 1 to 9 in case D, with a median effect of $1/2$, and effects that vary from $1/10$ at dose 1 to $9/10$ at dose 9. The dose-weighted signed rank statistic, T_{dose}, gives greater weight to pairs with larger doses, but in case A it equals Wilcoxon's signed rank statistic because there is only one dose. With the median dose fixed, there is some advantage to having several dose levels, but the advantage is not extremely large.

In cases E through G, there is a single dose, with a slightly larger median effect. A single dose with a median effect of 0.60 is comparable in terms of design sensitivity to $L = 5$ doses with a median effect of 0.50. A single dose with a median effect of 0.75 or 0.83 in cases F and G produces larger design sensitivities $\widetilde{\Gamma}$ than are found in cases A–D. Faced with a choice between a single larger dose or a continuum of doses with a smaller median dose, the former would be preferred. This addresses one of the controversies in Sect. 5.2.5.

It is important to realize that cases E through G are entirely practical. An investigator in case B could, in effect, switch to case F by using only the high-dose pairs, those with dose $\delta_3 = 3$, because in that dose category, the median effect is $\beta\delta_3 = (1/4) \times 3 = 0.75$. In parallel, an investigator in case C could, in effect, switch to case G by using only the high-dose pairs, those with $\delta_3 = 5$, because in that dose category, the median effect is $\beta\delta_5 = (1/6) \times 5 = 0.83$. Discarding the low-dose pairs would decrease the sample size, but an increase in design sensitivity is often more important in a sensitivity analysis, as has been seen in various contexts, including Table 5.5 and Chap. 16.

Table 18.3 considers the design sensitivity that would result in cases B, C, and D of Table 18.2 if pairs with doses below the median were discarded, and T_{dose} were applied to the remaining pairs. The increases in design sensitivity, $\widetilde{\Gamma}$, are moderate for Normal errors, somewhat noticeable for logistic errors, and negligible for Cauchy errors.

---

possibilities has the same probability, namely $E(W_{ik}) = E(W_{ki})$ which equals Ψ if $d_i = d_k = 1$. This shows $\Psi = p_1^{'}/2$.

Table 18.2 Design sensitivity with doses

Case	L	β	$\beta\,(L+1)\,/2$	Normal	Logistic	Cauchy
A	1	1/2	0.50	3.2	2.0	1.8
B	3	1/4	0.50	3.8	2.2	2.0
C	5	1/6	0.50	4.0	2.2	2.1
D	9	1/10	0.50	4.2	2.3	2.1
E	1	3/5	0.60	4.0	2.2	2.0
F	1	3/4	0.75	5.9	2.7	2.4
G	1	5/6	0.83	7.4	3.1	2.6

There are L equally probable, equally spaced, integer doses, $\delta_1 = 1, \ldots, \delta_L = L$ and the effect is proportional to dose with slope β, so the median effect is $\beta(L+1)/2$. The table gives the design sensitivity for the signed rank statistic with doses, which reduces to Wilcoxon's signed rank statistic when there is only one dose, $L = 1$

Table 18.3 Design sensitivity with doses, discarding pairs with doses below the median dose

Case	L	β	$\beta\,(3L+1)\,/4$	Normal	Logistic	Cauchy
H	3	1/4	0.625	4.5	2.4	2.2
I	5	1/6	0.667	5.0	2.5	2.3
J	9	1/10	0.700	5.5	2.7	2.3

Initially, there are L equally probable, equally spaced, integer doses, with L odd, $\delta_1 = 1, \ldots, \delta_L = L$ and the effect is proportional to dose with slope β. Unlike Table 18.2, pairs with doses strictly below $(L+1)/2$ are discarded, so the median dose for the remaining pairs is $(3L+1)/4$ and the median effect in these pairs is $\beta(3L+1)/4$. The table gives the design sensitivity for the signed rank statistic with doses applied to the pairs that are not discarded

Tables 18.2 and 18.3 refer to a dose that affects the magnitude of the effect. Not all doses do this. In one context, years of exposure may be unimportant, and yesterday's dose may be all-important. In another context, the dose a decade ago may matter, but yesterday's dose may be inconsequential for today's response. Doses may be measured with substantial errors, so that an indirect use of doses is superior to their direct incorporation in a test statistic [16, 25]. The wrong dose may be measured or the right one may be measured poorly.

Arguably, Tables 18.2 and 18.3 speak not to the value of doses taken in isolation, but rather to the value of treated subjects who experienced only small effects when treatment effects are heterogeneous. When doses are actually indicative of the magnitude of the effect, they serve to identify treated subjects who will experience small effects. Here and in Chap. 17, design sensitivity is greater if little weight is given to subjects experiencing negligible effects.

18.5 Example: Maimonides' Rule

Recall from Sects. 1.3 and 5.2.3 the study by Angrist and Lavy [1] of class size manipulated by Maimonides' rule. So far, the discussion has considered only the 86 pairs in Fig. 1.1 in which a school with between 41 and 50 students in the fifth

Table 18.4 High, lower, and mixed-dose pairs in Angrist and Lavy's study of class size manipulated by Maimonides' rule

	86 high-dose pairs	49 lower-dose pairs	135 mixed-dose pairs
Class size	−10.88	−5.33	−7.38
Math test	3.97	3.47	3.66
Verbal test	2.98	1.45	2.55

The table describes matched pair differences, larger-cohort-minus-smaller-cohort. The tabled values are trimeans, a robust location estimate that is twice the median plus the quartiles divided by four

grade was paired with a school with between 31 and 40 students in fifth grade. These are the "high-dose" pairs, in that adherence to Maimonides' rule would cut a class of 40 students into two classes of average size 20.5 if one additional student enrolled. In Fig. 1.1, adherence to Maimonides' rule is imperfect, so the typical reduction in class size is smaller than 20 students. Table 18.4 considers the addition of 49 more pairs, for schools with between 70 and 90 students in the fifth grade. These are the "lower-dose" pairs, in that adherence to Maimonides' rule would cut two classes of 40 students into three classes of average size 27 if one additional student enrolled. Table 18.4 gives the typical difference in class size and test scores in matched pairs, larger-cohort-minus-smaller-cohort, for the 86 "high-dose" pairs, the 49 "lower-dose" pairs, and the 135 "mixed-dose" pairs formed by pooling the pairs, $135 = 86 + 49$. In Table 18.4, the typical difference in class size is larger for the "high-dose" pairs, −10.9 students, than for the "lower-dose" pairs, −5.3 students, and the difference in verbal test scores is also larger.

Table 18.5 compares the sensitivity analysis in Table 5.3 with two additional sensitivity analyses. One analysis uses Wilcoxon's signed rank statistic and the coherent signed rank statistic applied to the 135 "mixed-dose" pairs. The other analysis uses the dose-weighted signed rank statistic and the coherent version of that statistic [12] applied to the 135 "mixed-dose" pairs with doses of 2 and 1 for "high-dose" and "lower-dose" pairs, respectively. In this one example, the coherent statistic using the 86 "high-dose" pairs is marginally less sensitive than the other analyses to departures from random treatment assignment. One should not make too much of individual examples, such as Tables 5.5 and 18.5; however, the results in those tables are not incompatible with the theoretical results about design sensitivity in this chapter.

18.6 Reactive Doses

In most randomized experiments, the dose of treatment is an aspect of the treatment inflicted upon an experimental subject. Such a dose is called "nonreactive." Nonreactive doses need to be distinguished from "reactive doses" and "purely reactive

Table 18.5 Sensitivity analysis and coherence in Angrist and Lavy's study of academic test performance and class size manipulated by Maimonides' rule

	86 high-dose pairs			135 mixed-dose pairs			135 dose-weighted pairs		
Γ	Math	Verbal	Coherent	Math	Verbal	Coherent	Math	Verbal	Coherent
1.00	0.0012	0.00037	0.00018	0.00034	0.00023	0.000081	0.00034	0.00014	0.000050
1.45	0.057	0.027	0.015	0.058	0.046	0.015	0.049	0.029	0.013
1.65	0.138	0.075	0.043	0.170	0.142	0.078	0.145	0.096	0.051

The table gives upper bounds on the one-sided P-values. The sensitivity analysis from Table 5.3 in Sect. 5.2.3 is the analysis for the 86 "high-dose" pairs. The remaining analyses add 49 "lower-dose" pairs to make 135 "mixed-dose" pairs. One analysis uses the signed rank statistic giving all pairs equal weights. The other analysis gives the "high-dose" pairs twice the weight of the "lower-dose" pairs. At $\Gamma = 1.65$, the only upper bound that is <0.05 is for coherence with the 86 "high-dose" pairs

doses" that rarely occur in experiments but may occur in observational studies [6, 19].

In an observational study, the "dose" of treatment may have an ambiguous status, because the investigator did not control the dose; rather, the investigator observes a number and claims that it is a dose. A dose is "reactive" if its recorded value reflects both an aspect of the treatment inflicted upon an individual and also the individual's reaction to that treatment. Biomarkers of exposure to toxins, such as cotinine as a measure of exposure to tobacco, reflect a combination of the intensity of exposure to the toxin and the individual's reaction to the toxin [8]. It is entirely possible that genetically different people exposed to the same dose of a toxin will differ in their levels of a biomarker measuring a metabolic response to the toxin, and it is entirely possible that these genetic differences will become confused with effects caused by the toxin [22]. For this reason, Weisskopf and Webster [28] prefer a coarse but nonreactive dose to a possibly more precise but also possibly reactive dose. See also [6].

At the opposite end of the spectrum from a nonreactive dose is a purely reactive dose. A purely reactive dose is entirely unaffected by the specific treatment under study; that is, such a "dose" takes the same value whether an individual is assigned to treatment or control, but it varies from one individual to the next. A purely reactive dose may truly reflect a dose of something, but it is not a measure of the intensity of the specific treatment under study. To claim that a dose is purely reactive is to claim that associations of outcomes with that dose are spurious as indicators of the effects caused by the treatment under study. Consider a study of some version of smoking cessation counselling and its effect on a health outcome, such as lung function. Suppose, further, that this version of smoking cessation counselling is ineffective, that it does not alter the amount smoked. Cotinine may truly reflect unmeasured tobacco exposure, but both cotinine and tobacco exposure may both be unaffected by this version of smoking cessation counselling. If heavy smokers avoid smoking cessation counselling, then cotinine levels may be lower among those counselled, and lung function may be correspondingly better, but these lower cotinine levels do not reflect the intensity or effectiveness of the counselling; rather, they reflect a selection bias in who receives counselling. This is true even if cotinine levels do reflect tobacco exposure, and by virtue of reflecting tobacco exposure, do predict health outcomes.

Expressed differently, a nonreactive dose is an aspect of the treatment inflicted on an individual. A reactive dose is an outcome that is altered by the treatment inflicted upon an individual. A purely reactive dose is an unaffected outcome, that is, an outcome that is not altered by the treatment inflicted upon an individual.

The dose-weighted signed rank statistic [12, 14, 27] of Sect. 18.4, namely $T_{\text{dose}} = \sum_{i=1}^{I} \text{sgn}\,(Y_i)\; q_i\, d_i$, is applicable only if the dose d_i is nonreactive. Alternative analyses along with their design sensitivities are discussed in [19].

18.7 Further Reading

Coherence and doses were discussed in Chap. 5, where general references may be found. The relationship between design sensitivity and coherence is discussed in [4, 15, 17, 20, 21]. The coherent signed rank statistic was proposed in [12]. General versions of the methods in Sect. 18.3 for discovering coherence are discussed in [20, 21]. These general methods use the m-statistics of Sect. 2.8 rather than rank tests, and they do not require matched pairs. An alternative to considering infinitely many test statistics in Scheffé's method is to limit attention to two specific statistics, with a smaller correction for multiple testing, a method known as testing twice; see Sect. 19.3 and [18]. The relationship between design sensitivity and doses is discussed in [15, 17]. Related issues are discussed in [12, 14] and [13, §9]. Reactive doses, their analysis, and related aspects of design sensitivity are discussed in [19]; see also [6]. An alternative formulation of coherence is given by K. Joreskog and Arthur Goldberger [5]. Dose errors are discussed in [16, 25]. Christopher Wild [29] surveys the many types of doses that arise in modern cancer epidemiology. The proof in the appendix is from [17].

18.8 Software

The coherent signed rank statistic is implemented in the function `cohere` in the `DOS2` package in `R`. The methods in Sect. 18.3, including Scheffé's method, are implemented by `cohere` for the coherent signed rank statistic, and for a broader class of statistics in the `sensitivitymult` package. This broader class of statistics includes means and other m-statistics.

18.9 Data

The data from Angrist and Lavy's [1] is in `angristlavy` in the `DOS2` package, and the help-file for `cohere` reproduces several of the analyses done in this chapter. The `DOS2` package also includes an observational study `teeth` from [20] with a bivariate outcome. The study concerns the effects of smoking on periodontal disease, where the effects may be different for upper and lower teeth. The help-file for `cohere` applies the methods of this chapter to `teeth`.

Appendix: Proof of Proposition 18.1

Proof Define $\Psi_1 = E(d_i W_{ii})$ and $\Lambda_1 = E(d_i V_{ii}) = E(d_i)$. In the favorable situation in the statement of Proposition 18.1, the expectation of T_{dose} is

$$E(T_{\text{dose}}) = E\left(\sum_{i=1}^{I}\sum_{k=1}^{I} d_i W_{ik}\right) = \sum_{i=1}^{I} E(d_i W_{ii}) + \sum_{i=1}^{I}\sum_{k\neq i} E(d_i W_{ik})$$
$$= I\Psi_1 + I(I-1)\Psi.$$

In the sensitivity analysis for T_{dose}, the upper bound on the one-sided P-value is obtained by comparing T_{dose} to the distribution of the sum, $\overline{\overline{T}}_{\Gamma,\text{dose}}$, of I independent random variables, $i = 1, \ldots, I$, taking the value $d_i\, q_i$ with probability $\theta = \Gamma/(1+\Gamma)$ and the value 0 with probability $1-\theta$, so $E\left(\overline{\overline{T}}_{\Gamma,\text{dose}} \,\middle|\, \mathscr{F}\right) = \theta \sum_{i=1}^{I} d_i\, q_i$; see [12, 14].Because $d_i\, q_i = \sum_{k=1}^{I} d_i V_{ik}$,

$$E\left(\overline{\overline{T}}_{\Gamma,\text{dose}} \,\middle|\, \mathscr{F}, \mathscr{Z}\right) = \theta \sum_{i=1}^{I} d_i\, q_i = \theta \sum_{i=1}^{I}\sum_{k=1}^{I} d_i V_{ik},$$

so that

$$E\left(\overline{\overline{T}}_{\Gamma,\text{dose}}\right) = E\left\{E\left(\overline{\overline{T}}_{\Gamma} \,\middle|\, \mathscr{F}, \mathscr{Z}\right)\right\} = \theta \sum_{i=1}^{I}\sum_{k=1}^{I} E(d_i V_{ik})$$
$$= \theta\left\{I\Lambda_1 + I(I-1)\Lambda\right\}.$$

Using the result in [15, §3], the design sensitivity, $\widetilde{\Gamma}$, is the limit, as $I \to \infty$, of the solutions to $E(T_{\text{dose}}) = E\left(\overline{\overline{T}}_{\Gamma,\text{dose}}\right)$, or equivalently to

$$\frac{\Gamma}{1+\Gamma} = \frac{I\Psi_1 + I(I-1)\Psi}{I\Lambda_1 + I(I-1)\Lambda}, \tag{18.14}$$

where the right side of (18.14) tends to Ψ/Λ as $I \to \infty$, yielding (18.13).

Bibliography

1. Angrist, J.D., Lavy, V.: Using Maimonides' rule to estimate the effect of class size on scholastic achievement. Q. J. Econ. **114**, 533–575 (1999)
2. Caughey, D., Dafoe, A., Seawright, J.: Nonparametric combination (NPC): a framework for testing elaborate theories. J. Polit. **79**, 688–701 (2017)

3. Cox, D.R.: A note on data-splitting for the evaluation of significance levels. Biometrika **62**, 441–444 (1975)
4. Heller, R., Rosenbaum, P.R., Small, D.: Split samples and design sensitivity in observational studies. J. Am. Stat. Assoc. **104**, 1090–1101 (2009)
5. Joreskog, K.G., Goldberger, A.S.: Estimation of a model with multiple indicators and multiple causes of a single latent variable. J. Am. Stat. Assoc. **70**, 631–639 (1975)
6. Karmakar, B., Small, D.S., Rosenbaum, P.R.: Using evidence factors to clarify exposure biomarkers. Am. J. Epidemiol. **189**, 243–249 (2020)
7. Noether, G.E.: Elements of Nonparametric Statistics. Wiley, New York (1967)
8. Perera, F.P., Weinstein, I.B.: Molecular epidemiology: recent advances and future directions. Carcinogenesis **21**, 517–524 (2000)
9. Rao, C.R.: Linear Statistical Inference and Applications. Wiley, New York (1973)
10. Rosenbaum, P.R.: From association to causation in observational studies. J. Am. Stat. Assoc. **79**, 41–48 (1984)
11. Rosenbaum, P.R.: The role of known effects in observational studies. Biometrics **45**, 557–569 (1989)
12. Rosenbaum, P.R.: Signed rank statistics for coherent predictions. Biometrics **53**, 556–566 (1997)
13. Rosenbaum, P.R.: Observational Studies, 2nd edn. Springer, New York (2002)
14. Rosenbaum, P.R.: Does a dose-response relationship reduce sensitivity to hidden bias? Biostatistics **4**, 1–10 (2003)
15. Rosenbaum, P.R.: Design sensitivity in observational studies. Biometrika **91**, 153–164 (2004)
16. Rosenbaum, P.R.: Exact, nonparametric inference when doses are measured with random errors. J. Am. Stat. Assoc. **100**, 511–518 (2005)
17. Rosenbaum, P.R.: What aspects of the design of an observational study affect its sensitivity to bias from covariates that were not observed? In: Dorans, N.J., Sinharay, S. (eds.) Looking Back: Proceedings of a Conference in Honor of Paul W. Holland. Lecture Notes in Statistics Book, vol. 202, pp. 87–114. Springer, New York (2011)
18. Rosenbaum, P.R.: Testing one hypothesis twice in observational studies. Biometrika **99**, 763–774 (2012)
19. Rosenbaum, P.R.: Nonreactive and purely reactive doses in observational studies. In: Berzuini, C., Dawid, A.P., Bernardinelli, L. (eds.) Causality: Statistical Perspectives and Applications, pp. 273–289. Wiley, New York (2012)
20. Rosenbaum, P.R.: Using Scheffé projections for multiple outcomes in an observational study of smoking and periodontal disease. Ann. Appl. Stat. **10**, 1147–1471 (2016)
21. Rosenbaum, P.R.: Combining planned and discovered comparisons in observational studies. Biostatistics **21**, 384–399 (2020). https://doi.org/10.1093/biostatistics/kxy055
22. Savitz, D.A., Wellenius, G.A.: Exposure biomarkers indicate more than just exposure. Am. J. Epidemiol. **187**, 803–805 (2017)
23. Scheffé, H.: A method for judging all contrasts in the analysis of variance. Biometrika **40**, 87–104 (1953)
24. Small, D., Rosenbaum, P.R.: War and wages: the strength of instrumental variables and their sensitivity to unobserved biases. J. Am. Stat. Assoc. **103**, 924–933 (2008)
25. Small, D.S., Rosenbaum, P.R.: Error-free milestones in error-prone measurements. Ann. Appl. Stat. **3**, 881–901 (2009)
26. Tchetgen Tchetgen, E.J.: The control outcome calibration approach for causal inference with unobserved confounding. Am. J. Epidemiol. **179**, 633–640 (2013)
27. van Eeden, C.: An analogue, for signed rank statistics, of Jureckova's asymptotic linearity theorem for rank statistics. Ann. Math. Stat. **43**, 791–802 (1972)
28. Weisskopf, M.G., Webster, T.F.: Trade-offs of personal versus more proxy exposure measures in environmental epidemiology. Epidemiology **28**, 635–643 (2017)
29. Wild, C.P.: Environmental exposure measurement in cancer epidemiology. Mutagenesis **24**, 117–125 (2009)

30. Zhang, K., Small, D.S., Lorch, S., Srinivas, S., Rosenbaum, P.R.: Using split samples and evidence factors in an observational study of neonatal outcomes. J. Am. Stat. Assoc. **106**, 511–524 (2011)
31. Zhao, Q., Small, D.S., Rosenbaum, P.R.: Cross-screening in observational studies that test many hypotheses. J. Am. Stat. Assoc. **113**, 1070–1084 (2018)

Chapter 19
Choice of Test Statistic

Abstract The choice of one test statistic rather than another affects the design sensitivity, as it affects the power and efficiency of a randomization test. This is demonstrated by computing the design sensitivity for several competing test statistics in the same sampling situation. Familiar test statistics, such as Wilcoxon's signed rank statistic and the mean of the matched pair differences, are often inferior to test statistics constructed to have larger design sensitivities. There is no one uniformly best test statistic in terms of design sensitivity, raising the possibility of adaptive inference, in which the data guide the choice of test statistic. Bahadur efficiency provides a comparison of the performance of two test statistics when used in a sensitivity analysis at a value of Γ below the minimum of their two design sensitivities.

19.1 Choice of Test Statistic Affects Design Sensitivity

19.1.1 Design Anticipates Analysis

Design anticipates analysis. A more effective research design produces data that will be more decisive when analyzed. An effective design must anticipate an effective analysis. But what makes an analysis effective?

The investigator planning an observational study hopes to be in a position to say that the substantive conclusions are insensitive to small and moderate biases from unmeasured covariates. Previous chapters examined elements of the research design that affect the design sensitivity.

As it turns out, the choice of analytical techniques also affects the design sensitivity. This is not surprising. The choice of analytical techniques affects the power and efficiency of a randomized trial. A better choice of test statistic in a randomized trial has a higher probability of rejecting a false null hypothesis, thereby excluding that hypothesis from a confidence interval, thereby producing a shorter, more informative confidence interval. Why shouldn't the choice of analytical techniques also affect the results of a sensitivity analysis and the design sensitivity? Indeed, they do. Indeed, we have seen this before.

© Springer Nature Switzerland AG 2020

P. R. Rosenbaum, *Design of Observational Studies*, Springer Series in Statistics,
https://doi.org/10.1007/978-3-030-46405-9_19

Chapter 17 considered situations in which only a subset of the population "responds" to treatment, the rest being unaffected [6]. In such situations, Proposition 17.1 and Table 17.3 found that the choice of test statistic had substantial consequences for the design sensitivity $\widetilde{\Gamma}$: Wilcoxon's signed statistic had much lower design sensitivity than Stephenson's [43] statistic. With a subpopulation of responders, Wilcoxon's statistic will, in large samples, exaggerate the degree to which small unobserved biases can explain away an ostensible treatment effect, but Stephenson's [43] statistic will not do this. Wilcoxon's statistic has some attractive properties in a randomized experiment when a treatment has about the same small effect on everyone, but it is not the best method of analysis when only a small subpopulation exhibits a large response to treatment. The average treatment effect may be small if most people are unaffected by treatment, but the benefits for the subpopulation of responders may be large.

Is the situation in Chap. 17 an unusual one? Or does the choice of test statistic typically play an important role in determining the conclusion of a sensitivity analysis?

19.1.2 A Simple Case: Step-Rank Statistics, Additive Effects, Normal Errors

Suppose that we are in the favorable situation in an observational study, with a genuine treatment effect and no bias from unmeasured covariates. If we were in the favorable situation, we would not know it. In this situation, we hope to report that the ostensible treatment effect cannot be explained away by small or moderate biases. The design sensitivity $\widetilde{\Gamma}$ is a guide: in very large samples, we are likely to report that rejection of the null hypothesis of no effect is insensitive to a bias $\Gamma < \widetilde{\Gamma}$, and we are likely to report that rejection is sensitive to a bias $\Gamma > \widetilde{\Gamma}$.

To simplify further, suppose that the favorable situation has I treated-minus-control matched pair differences, Y_i, $i = 1, \ldots, I$, drawn independently from the Normal distribution with expectation τ and variance 1, so the average treatment effect is τ. A claim that the ostensible treatment effect is τ must address the inevitable counterclaim that biased treatment assignment could produce the misleading appearance of a treatment effect.

In this situation in Chap. 15, with $\tau = 0.5$, the design sensitivity of Wilcoxon's signed rank statistic testing $H_0 : \tau = 0$ was found to be $\widetilde{\Gamma} = 3.17$. How does that compare to other statistics in the statistical literature?

Some statistics are much worse. The sign test statistic counts the number of positive Y_i and compares that count to the binomial distribution with I independent flips of a fair coin, a coin with probability 1/2 of a head. With $\tau = 0.5$, the sign test has design sensitivity $\widetilde{\Gamma} = 2.24$, far below $\widetilde{\Gamma} = 3.17$ for Wilcoxon's test [30, Table 2]. As $I \to \infty$, a sensitivity analysis performed with $\Gamma = 2.5$ will have power increasing to one if Wilcoxon's statistic is used, or power declining to zero

if the sign test is used. Perhaps that is not entirely surprising: for Normal data in a randomized experiment, Wilcoxon's test has much higher Pitman efficiency[1] than the sign test, so in this case Pitman efficiency and design sensitivity point in the same direction. Is that always true? It is not.

Bruce Brown [5] proposed a simple competitor of Wilcoxon's signed rank statistic. The statistic sorts the I pairs into increasing order of their absolute differences, $|Y_i|$. Brown ignores roughly $I/3$ pairs with the smallest $|Y_i|$. For the middle $I/3$ pairs, Brown counts 1 if Y_i is positive or 0 otherwise. For the largest $I/3$ pairs, Brown counts 2 if Y_i is positive or 0 otherwise. His test statistic is the total count. Under $H_0 : \tau = 0$, the null distribution of this statistic is that of a weighted sum of two binomial distributions. The same null distribution arises as a randomization distribution in a paired randomized experiment testing Fisher's hypothesis of no treatment effect. Brown [5, Table 1] shows that his statistic has good robustness properties, but is slightly less efficient than Wilcoxon's statistic for Normal data. For $\tau = 0.5$, Brown's statistic has design sensitivity $\widetilde{\Gamma} = 3.60$, noticeably higher than $\widetilde{\Gamma} = 3.17$ for Wilcoxon's statistic. The improved design sensitivity of Brown's statistic is a consequence of ignoring pairs with small $|Y_i|$.

Edward Markowski and Thomas Hettmansperger [21] considered a similar class of statistics, but did not divide the $|Y_i|$ into thirds; see also [8, 10]. Within this larger class, much higher design sensitivities occur. Consider a statistic that ignores the smallest $3/4 = 6/8$ of the $|Y_i|$, scores 1 for a positive Y_i in the next 1/8th, and scores 2 for a positive Y_i in the top 1/8th. In a randomized experiment, that statistic has poor Pitman efficiency, but it has design sensitivity of $\widetilde{\Gamma} = 6.55$, more than twice $\widetilde{\Gamma} = 3.17$ for Wilcoxon's statistic [30, Table 6]. The gains from this particular statistic are still present but smaller if the Y_i are logistic rather than Normal, and they disappear if the Y_i have a t-distribution with 3 degrees of freedom [30, Table 6]. The statistic just described is well-suited to Normal errors, producing a large increase in design sensitivity; however, other statistics in the same family produce a smaller but still large increase in design sensitivity for Normal errors, but maintain that increase for long-tailed error distributions, such as the t-distribution with 3 degrees of freedom.

Gottfried Noether [23] considered an even simpler test. He counts the number of positive Y_i among the $\eta \times I$ largest $|Y_i|$. Under the null hypothesis in a randomized experiment, this count is binomial with sample size $\eta \times I$ and probability of success 1/2. In a sensitivity analysis in an observational study, the upper bound on the one-sided P-value is obtained from a binomial distribution with sample size $\eta \times I$ and probability of success $\Gamma/(1 + \Gamma)$. Consider $\eta = 1/3$, so attention focuses on the 1/3 of matched pairs with the largest $|Y_i|$.[2] Also, set the additive effect τ to be 1/2

[1] Various forms of efficiency, including Pitman efficiency, are explained in Sect. 19.5. If Pitman efficiency is unfamiliar, then read "efficiency" informally as "making good use of data" until Sect. 19.5. Pitman efficiency is the most standard form of large-sample efficiency, or asymptotic efficiency; indeed, so familiar that unnamed efficiency is understood to refer to Pitman efficiency.

[2] A reader of Noether's paper [23] should take note that Noether adopts a different notation and advocates $\eta = 2/3$ rather than $\eta = 1/3$.

the standard deviation of the Y_i for the Normal, logistic, and t-distribution with 3 degrees of freedom. Then Noether's statistic has design sensitivities $\widetilde{\Gamma}$ of 4.97 for the Normal, 4.72 for the logistic, and 5.77 for the t-distribution with 3 degrees of freedom, while Brown's statistic has design sensitivities of 3.60 for the Normal, 3.83 for the logistic, and 5.39 for the t-distribution [30, Table 2]. Noether's statistic has much worse Pitman efficiency than Brown's statistic, but much higher design sensitivity, by virtue of ignoring 2/3 of the pairs with the smallest $|Y_i|$.

Briefly, to make this tangible, I sampled $I = 1000$ independent differences Y_i from a Normal distribution with expectation $\tau = 1/2$ and variance 1. So I is large, but not unrealistically so. Let us compute the upper bound on the one-sided P-value testing the null hypothesis of no treatment effect in the presence of a bias of at most $\Gamma = 3.2$. This will be done three times, using Wilcoxon's signed rank statistic, using Brown's [5], and using Noether's [23] with $\eta = 1/3$. For these three statistics in this sampling situation, the design sensitivities are, respectively, $\widetilde{\Gamma} = 3.17$, 3.60, and 4.97. As one might anticipate from their design sensitivities, at $\Gamma = 3.2$ the upper bound on the P-value was 0.387 for Wilcoxon's test, 0.081 for Brown's test, and 0.0087 for Noether's test. As $I \to \infty$, the P-value bound for Wilcoxon's test would tend to 1 because $\widetilde{\Gamma} = 3.17 < 3.2 = \Gamma$, but the P-value bounds for Brown's test and Noether's test would tend to 0 because in both cases $\widetilde{\Gamma} > 3.2 = \Gamma$. For the results of 10,000 repetitions of this experiment with varied sample sizes, see [30, Table 3].

This phenomenon is not difficult to understand. In studying the power of signed rank statistics, Willem Albers et al. [1] considered a function, $\text{abz}(y)$, defined for $y > 0$, as the probability that Y_i is positive conditionally given $|Y_i| = y$. The $\text{abz}(y)$ function tells you which $|Y_i|$ to ignore if the goal is a large design sensitivity, $\widetilde{\Gamma}$. The rule is simple: if $\text{abz}(y) > \Gamma/(1+\Gamma)$ then, at this value of y, positive Y_i occur too often to be explained away as a bias of magnitude Γ. You want to ignore the pairs $|Y_i| = y$ with $\text{abz}(y) < \Gamma/(1+\Gamma)$. For plots of $\text{abz}(y)$ for various distributions and further development, see [28].

19.2 Statistics Built for Superior Design Sensitivity

19.2.1 Building New Statistics for Use in Sensitivity Analyses

The families of statistics discussed in Sect. 19.1 were not built to increase design sensitivity; however, these large families contain particular statistics with substantially higher design sensitivity than that of Wilcoxon's test. Can statistics be built to perform well in sensitivity analyses? The statistics in Sect. 19.1 had larger design sensitivities than Wilcoxon's statistic, but lower Pitman efficiency in a randomization test, $\Gamma = 1$, for Normal errors. Does a gain in design sensitivity always entail a loss of efficiency in a randomization test? Is there a useful concept of efficiency for sensitivity analyses with $1 < \Gamma < \widetilde{\Gamma}$?

19.2.2 A New U-Statistic with Superior Design Sensitivity

Pick three integers, $(m, \underline{m}, \overline{m})$ with $1 \le \underline{m} \le \overline{m} \le m < I$. Suppose that we examine m of the I matched pair differences, Y_i, sort them into increasing order by $|Y_i|$, and count the number of positive Y_i in positions $\underline{m}, \underline{m}+1, \ldots, \overline{m}$, yielding an integer in $0, 1, \ldots, \overline{m} - \underline{m} + 1$. There are $I!/m!(I-m)!$ ways to pick m of I differences Y_i. Suppose that we do this calculation in all possible ways, and compute their average, calling the result T. The statistic T is an instance of a U-statistic, a class of statistics proposed by Wassily Hoeffding [13]. Erich Lehmann [17] provides an introduction to U-statistics, and Robert Serfling [41, §5] provides a detailed discussion.

In particular, Lehmann [17] observes that the statistic $(m, \underline{m}, \overline{m}) = (2, 2, 2)$ is almost identical to Wilcoxon's signed rank statistic. The statistic $(m, \underline{m}, \overline{m}) = (1, 1, 1)$ is the sign test statistic. The statistic $(m, \underline{m}, \overline{m}) = (m, m, m)$ is Stephenson's statistic [43], and it played an important role in Chap. 17. The general statistic $(m, \underline{m}, \overline{m})$ was introduced in [29] and further evaluated in [34]. Like Wilcoxon's statistic and Stephenson's statistics, the general statistic $(m, \underline{m}, \overline{m})$: (1) is a signed rank statistic, (2) changes gradually as the rank of $|Y_i|$ changes, unlike the group-rank statistics in Sect. 19.1, (3) has a simple formula [29, Proposition 1] for its design sensitivity, $\widetilde{\Gamma}$, (4) permits efficiency calculations [34] for $1 \le \Gamma < \widetilde{\Gamma}$.

Consider, for example, the statistic $(m, \underline{m}, \overline{m}) = (5, 4, 5)$. For the first $m = 5$ pairs, the statistic sorts and renumbers the 5 treated-minus-control pair differences in outcomes so that $|Y_1| < \cdots < |Y_5|$, then scores a 0 if $Y_4 < 0$ and $Y_5 < 0$, a 2 if $Y_4 > 0$ and $Y_5 > 0$, and otherwise scores a 1. This quantity is computed from every subset of $m = 5$ of the I pairs and averaged. Once averaged over the $I!/5!(I-5)!$ choices of $m = 5$ of the Y_i, the statistic has emphasized the signs of Y_i with larger $|Y_i|$, but it has totally ignored only the signs of the three smallest $|Y_i|$, unlike the statistics in Sect. 19.1 that ignored the signs of many of the $|Y_i|$. Although this description ignores the possibility of ties among the $|Y_i|$ and the possibility that some $Y_i = 0$, these possibilities merely require a more detailed discussion, but present no problems [29].

If the Y_i are independent with the same Normal distribution, with constant treatment effect τ and variance 1, then Wilcoxon's signed rank test and the statistic $(m, \underline{m}, \overline{m}) = (5, 4, 5)$ have, to two decimal places, Pitman relative efficiency 1.00 in a randomization test, $\Gamma = 1$, so this calculation provides no reason to prefer one test to the other; see [29, Table 1]. However, if the treatment effect is half the standard deviation of a pair difference, $\mathrm{E}(Y_i) = \tau = 1/2$, then the design sensitivity of Wilcoxon's test is $\widetilde{\Gamma} = 3.17$ but the design sensitivity of $(m, \underline{m}, \overline{m}) = (5, 4, 5)$ is higher $\widetilde{\Gamma} = 3.9$; see [29, Table 3].

What happens if the errors are long-tailed instead of Normal, having a t-distribution with 3 degrees of freedom? The t-distribution with 3 degrees of freedom has variance 3, hence standard deviation $\sqrt{3} = 1.732$, so consider an additive effect of $\tau = 1$ or $1/1.732 = 0.577$ of the standard deviation of a pair difference Y_i. In this case, Wilcoxon's statistic has design sensitivity $\widetilde{\Gamma} = 6.0$, but $(m, \underline{m}, \overline{m}) = (5, 4, 5)$ has higher design sensitivity $\widetilde{\Gamma} = 6.8$; see [29, Table 3].

Are there better choices than $(m, \underline{m}, \overline{m}) = (5, 4, 5)$? The statistic $(m, \underline{m}, \overline{m}) = (8, 6, 8)$ has Pitman efficiency of 0.98 relative to Wilcoxon's statistic with Normal errors in a randomization test, $\Gamma = 1$, so it is just slightly worse than Wilcoxon's test by that standard. However, with $\tau = 1/2$ and Normal errors, the design sensitivity of $(m, \underline{m}, \overline{m}) = (8, 6, 8)$ is $\widetilde{\Gamma} = 4.2$, and the design sensitivity of $(m, \underline{m}, \overline{m}) = (8, 7, 8)$ is $\widetilde{\Gamma} = 5.1$. In the long-tailed t-distribution situation of the previous paragraph, the design sensitivity of $(m, \underline{m}, \overline{m}) = (8, 6, 8)$ is $\widetilde{\Gamma} = 7.1$, while the design sensitivity of $(m, \underline{m}, \overline{m}) = (8, 7, 8)$ is $\widetilde{\Gamma} = 6.8$. See [29, Tables 1 and 3] for additional comparisons. So $(m, \underline{m}, \overline{m}) = (8, 7, 8)$ has larger design sensitivity than $(m, \underline{m}, \overline{m}) = (8, 6, 8)$ for Normal errors, but lower design sensitivity for errors from the t-distribution with 3 degrees of freedom; however, they are both better than Wilcoxon's statistic in both situations.

Randomization tests and sensitivity analyses invert to yield confidence intervals and point estimates for an additive treatment effect, τ; see Sects. 2.4 and 3.5. This is true for Wilcoxon's statistic, but also for the U-statistic $(m, \underline{m}, \overline{m})$; see [29, §5.1]. In a sensitivity analysis with fixed $\Gamma > 1$, as the sample size increases, $I \to \infty$, the confidence interval and the interval of point estimates converge to a real interval with nonzero length; see Sect. 16.3. If $1 < \Gamma < \widetilde{\Gamma}$, then that real interval for τ excludes zero. So, in a sensitivity analysis, as $I \to \infty$, a statistic such as $(m, \underline{m}, \overline{m}) = (8, 6, 8)$ with a larger design sensitivity will provide more informative point estimates and confidence intervals than a statistic with a smaller design sensitivity, such as Wilcoxon's statistic.

19.2.3 Example: Chemotherapy Associated Toxicity

Chapter 8 compared the treatment provided by medical oncologists (MOs) and gynecological oncologists (GOs) in $I = 344$ matched pairs of two similar women with ovarian cancer. Recall that MOs used chemotherapy more intensively than GOs, but survival was virtually the same, and chemotherapy-related toxicity was higher.

Specifically, recall Table 8.4, years 1–5, where the difference in toxicity was highly significant when Wilcoxon's signed rank test was compared to its randomization distribution, $\Gamma = 1$. At $\Gamma = 1.5$, the upper bound on the one-sided P-value is still small, 0.009, but a bias in treatment assignment of $\Gamma = 1.7$ would be large enough to lead to acceptance of no treatment effect at the conventional level of $\alpha = 0.05$ because the upper bound on the P-value is 0.080.

In fact, Wilcoxon's statistic has exaggerated the sensitivity to unmeasured bias. At $\Gamma = 1.7$, the upper bound on the P-value is 0.004 if the statistic $(m, \underline{m}, \overline{m}) = (8, 6, 8)$ is used, so a bias of $\Gamma = 1.7$ is not large enough to render plausible the hypothesis of no effect. Using (8,6,8), the upper bound on the P-value is 0.042 at $\Gamma = 2$. For these and additional calculations and examples, see [29, Table 2].

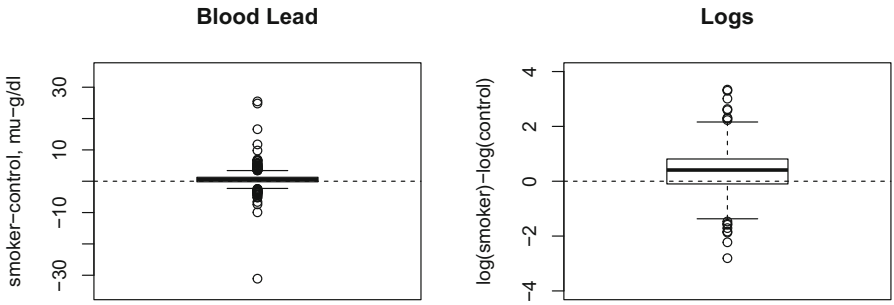

Fig. 19.1 Blood lead levels in daily cigarette smokers and matched never-smokers. The boxplot on the left is the difference in blood lead levels in μg/dl. The boxplot on the right is the difference of the natural logs of the lead levels. The dashed horizontal line is at zero difference

19.2.4 Example: Lead in the Blood of Smokers

Cigarette smoking is often found to be associated with elevated levels of lead in the blood [42]. Figure 19.1 depicts blood lead levels of 671 matched pairs of a daily cigarette smoker and a never smoker, matched for age, gender, education, income, and ethnicity. Smokers smoked every day for the past 30 days, and smoked at least 10 cigarettes per day. Never-smokers smoked fewer than 100 cigarettes in their lives and did not smoke at all in the past 30 days. The data are from the 2007–2008 US National Health and Nutrition Examination Survey and were used as an example in [31]. Blood lead is recorded as μg/dl. The left boxplot in Fig. 19.1 is the difference in lead levels; the right boxplot is the difference in the logs of lead levels.

Consider three sensitivity analyses for a test of no effect applied to the difference of logs of lead levels. The first analysis uses Wilcoxon's signed rank test, the second Brown's [5] statistic in Sect. 19.1, and the third uses the U-statistic $(m, \underline{m}, \overline{m}) = (8, 7, 8)$. All three analyses are evaluated at $\Gamma = 3.4$. The upper bound on the P-value is 0.324 for Wilcoxon's statistic, 0.048 for Brown's statistic, and 0.009 for $(m, \underline{m}, \overline{m}) = (8, 7, 8)$; see [31, Table 2]. So, once again, the choice of test statistic has consequences for the reported degree of sensitivity to unmeasured biases, with a pattern anticipated by the design sensitivity.

19.2.5 m-Statistics

M-statistics are the quantities that are equated to zero in defining Peter Huber's m-estimates [14]. Maritz [19, 20] developed the exact randomization distribution of m-statistics for matched pairs, and minor adjustments yield the randomization distribution of m-statistics in matched sets with one or more controls and their associated sensitivity analyses [26]. M-statistics and their randomization distributions were discussed and illustrated in Sect. 2.8.

Consider inference about an additive treatment effect, τ, with treated-minus-control matched pair differences Y_i. As discussed in Sect. 2.8, an m-statistic testing $H_0 : \tau = \tau_0$ is defined by its ψ-function and by a scale factor s_{τ_0}, most commonly $s_{\tau_0} = \text{median}(|Y_i - \tau_0|)$. In the discussion here, the ψ-function is monotone increasing and odd; that is, $y < y'$ implies $\psi(y) \leq \psi(y')$ and $\psi(y) = -\psi(-y)$. The m-statistic is then $T = \sum_{i=1}^{I} \text{sign}(Y_i - \tau_0)\psi(|Y_i - \tau_0|/s_{\tau_0})$ where $\text{sign}(y)$ is 1 if $y > 0$, 0 if $y = 0$, and -1 if $y < 0$.

Taking $\psi(y) = y$ yields the same inferences as the permutation distribution of the sample mean, $\overline{Y} = I^{-1} \sum_{i=1}^{I} Y_i$, the so-called permutational t-test. Huber [15] advocated $\psi_{\text{hu}}(y) = \max\{-\omega, \min(\omega, y)\}$ for some $\omega > 0$, and Sect. 2.8 illustrated $\omega = 1$.

There is a simple formula for the design sensitivity of an m-statistic in matched pairs [32, Corollary 1]. If the Y_i are independent Normal observations with expectation $\tau = 1/2$ and variance 1, then the design sensitivity of the permutational t-test is $\widetilde{\Gamma} = 3.5$, and the design sensitivity of Huber's $\psi_{\text{hu}}(y)$ with $\omega = 2$ is $\widetilde{\Gamma} = 3.3$; see [32, Table 3]. Recall that Wilcoxon's signed rank statistic has $\widetilde{\Gamma} = 3.17$ in this same situation, but the U-statistic $(m, \underline{m}, \overline{m}) = (8, 6, 8)$ has $\widetilde{\Gamma} = 4.2$.

The design sensitivity of m-statistics can be increased by inner trimming. Let ι be a number that defines the extent of inner trimming, with $0 \leq \iota < \omega$. The odd function $\psi_{\text{in}}(y)$ is zero for $-\iota \leq y \leq \iota$, is ω for $y \geq \omega$, rises linearly from 0 to ω on the interval $\iota \leq y \leq \omega$, and of course satisfies $\psi_{\text{in}}(y) = -\psi_{\text{in}}(-y)$ because it is required to be odd. Here, $\psi_{\text{in}}(y)$ ignores y near zero. If the Y_i are Normal with expectation $\tau = 1/2$ and variance 1, then the m-statistic with $\psi_{\text{in}}(y)$ defined by $\iota = 0.5$ and $\omega = 2$ has design sensitivity $\widetilde{\Gamma} = 4.0$, substantially higher than the permutational t-test.

As in the discussion of U-statistics, instead of Normal Y_i's, consider long-tailed errors having a t-distribution with 3 degrees of freedom, and an additive effect of $\tau = 1$, so the effect is 0.577 of the standard deviation of a pair difference Y_i. With $\iota = 0.5$ and $\omega = 2$, the design sensitivities $\widetilde{\Gamma}$ are: 5.6 for the permutational t-test, 6.0 for $\psi_{\text{hu}}(\cdot)$, and 7.0 for $\psi_{\text{in}}(\cdot)$. For the quoted design sensitivities of m-statistics, see [32, Table 3].

19.2.6 Summary

So far, in this chapter, we have seen that the reported degree of sensitivity to unmeasured bias depends not only on the sampling situation—the state of the world—but also on the investigator's choice of test statistic. More precisely, in the same sampling situation, two statistics may have very different design sensitivities,

$\widetilde{\Gamma}$, so that in sufficiently large samples the statistic with the larger design sensitivity is almost certain to report greater insensitivity to unmeasured bias.[3]

Statistics with superior design sensitivity give little importance to Y_i near zero, whether positive or negative. There are robust statistics that do this within the family of step-rank statistics in Sect. 19.1, within the family of U-statistics, and within the family of m-statistics. For some statistics, an increase in design sensitivity comes with a decrease in Pitman efficiency in a randomization test, $\Gamma = 1$; however, there are other statistics that perform well in both respects.

19.3 Adaptive Inference

19.3.1 Using the Data to Choose a Test Statistic

Although Wilcoxon's test often has lower design sensitivity than other tests, there is no one test that is always best. In large samples, the best test when the errors have short-tailed Normal distributions will not be the best test when the errors have long tails. For instance, the U-statistic $(m, \underline{m}, \overline{m}) = (8, 7, 8)$ performs extremely well with Normal errors, but $(m, \underline{m}, \overline{m}) = (8, 6, 7)$ down-weights outliers and is much better with long-tailed errors [29, Table 3].

Suppose that we are considering two competitive tests, T_1 and T_2, knowing that the design sensitivity of T_1 is higher than that of T_2 in some situations and lower in others. Is there some way that we can always use the better test, the one with the larger design sensitivity? In finite samples, the answer is no. However, we can make an adaptive choice between T_1 and T_2—a choice informed by the data—in such a way that as $I \to \infty$ our adaptive procedure always has the larger of the two design sensitivities of the two component tests, T_1 and T_2; see [30, 31, 40].

What is at stake here? If you pick the wrong statistic, the one with a smaller design sensitivity, then you will exaggerate the ease with which your findings may be dismissed as a bias in treatment assignment rather than an effect caused by the treatment. Moreover, this exaggeration of the weakness of your findings will persist as the sample size increases, $I \to \infty$. In this sense, much more is at stake in an

[3] This can be stated and demonstrated precisely. In a favorable situation with a treatment effect and no unmeasured biases, suppose test statistics T_1 and T_2 have design sensitivities $\widetilde{\Gamma}_1$ and $\widetilde{\Gamma}_2$, respectively, where $\widetilde{\Gamma}_1 > \widetilde{\Gamma}_2$. As judged by just one standard, namely design sensitivity, T_1 is the better statistic with the larger design sensitivity; however, as we have seen, T_1 might be inferior by other standards, such as Pitman efficiency in a randomization test, $\Gamma = 1$. Perform the sensitivity analysis at Γ where $\widetilde{\Gamma}_1 > \Gamma > \widetilde{\Gamma}_2$. Let $\mathscr{E}_I$ be the event that the sensitivity analysis at Γ rejects no effect at level α using T_2 but accepts using T_1. If event $\mathscr{E}_I$ occurs, then design sensitivity has misled you, encouraging you to use T_1 when only T_2 leads to rejection of the null hypothesis of no effect in the presence of a bias of Γ. It is not difficult to show that the probability of the event $\mathscr{E}_I$ declines to 0 very rapidly as the sample size increases, $I \to \infty$, specifically at an exponential rate [34, Proposition 3].

observational study than in a randomized experiment. In a randomized experiment, if you choose the less efficient of two consistent tests, then you will waste some of the sample by virtue of your poor choice, but if the experiment is made sufficiently large, $I \to \infty$, then you will sort things out correctly eventually. In contrast, in an observational study as $I \to \infty$, if Y_i is Normal with expectation $\tau = 1/2$ and variance 1, and if you choose to use Wilcoxon's test, then you will ultimately report that your study is sensitive to a bias of $\Gamma = 3.2$, where you could have reported insensitivity to a bias of $\Gamma = 4.9$ simply by switching to Noether's [23] test with $\eta = 1/3$. Surely, one can do better. As $I \to \infty$, it will eventually be clear if Y_i is not Normal but rather long-tailed, so we should be able to pick the better of two tests eventually, for large enough I. In Sect. 22.1, general methods are considered that split the data at random into two parts, one for planning the study, the other for analysis and conclusions [12, 48]. These split sample methods sacrifice part of the sample, but they make correct choices concerning design sensitivity $\widetilde{\Gamma}$ for sufficiently large I, so it is clear that we need not persist, as $I \to \infty$, in the use of T_1 if its design sensitivity is inferior to that of T_2. Adaptive methods take a different path to the same goal: they avoid sample splitting, perform one analysis, not two; yet, they eventually use the statistic with superior design sensitivity.

How does an adaptive procedure work? We want to test the null hypothesis H_0 of no treatment effect at level α, conventionally $\alpha = 0.05$, in the presence of a bias in treatment assignment of at most Γ, so assume H_0 is true and the bias is at most Γ. Stated informally, we determine the best joint critical values, $t_{\Gamma 1}$ and $t_{\Gamma 2}$ that equitably share α between the two tests such that

$$\Pr\left(T_1 \geq t_{\Gamma 1} \text{ or } T_2 \geq t_{\Gamma 2} \mid \mathscr{F}, \mathscr{Z}\right) \leq \alpha \tag{19.1}$$

while minimizing

$$\left|\Pr\left(T_1 \geq t_{\Gamma 1} \mid \mathscr{F}, \mathscr{Z}\right) - \Pr\left(T_2 \geq t_{\Gamma 2} \mid \mathscr{F}, \mathscr{Z}\right)\right|. \tag{19.2}$$

The minimization of (19.2) means that the tests share α equitably. We could achieve (19.1) using the Bonferroni inequality, using each test at level $\alpha/2$, but this would be wasteful. After all, T_1 and T_2 are two very similar tests of the same null hypothesis computed from the same data, so they are highly correlated, and use of the Bonferroni inequality would be quite conservative [31, Table 4]. Because the joint distribution of T_1 and T_2 is used in (19.1), the adaptive test is not conservative and pays only a small price for looking at two tests. Additionally, the adaptive test has the larger design sensitivity of the two component tests.

If we knew which test had the larger design sensitivity in the situation that generated our data, then of course we would use that test; however, with competitive tests, we do not know this. In simulations with $I = 100$ or $I = 250$ or $I = 500$ pairs, the power of the adaptive test at $\Gamma = 2$ or $\Gamma = 3$ is close to but slightly below the power of the better test; however, the power of the adaptive test is higher than, often dramatically higher than, the power of the inferior test [30, Table 3]. Indeed, if Γ is strictly between the design sensitivities of the two component tests, then the

power of the inferior test is tending to 0 as $I \to \infty$, while the power of the better test and the power of the adaptive test are both tending to 1.

19.3.2 Example: Adaptive Inference for Chemotherapy Associated Toxicity

Consider again the ovarian cancer example from Chap. 8 that was reanalyzed in Sect. 19.2. Suppose that we adapt between two statistics from Sect. 19.1, namely Brown's statistic and Noether's statistic with $\eta = 1/3$. Here, Noether's statistic, T_1, focuses on the 1/3 of the Y_i with the largest $|Y_i|$, while Brown's statistic, T_2, is emphasizing those same pairs, but is also giving half as much weight to the middle third of the $|Y_i|$. In this case, exact values of $t_{\Gamma 1}$ and $t_{\Gamma 2}$ in (19.1) are found by elementary manipulations of the two binomial distributions [30]. The adaptive test rejects at level $\alpha = 0.05$ the hypothesis of no effect of MO-versus-GO on chemotherapy-related toxicity in the presence of a bias of at most $\Gamma = 2.2$, and it does this using Noether's test rather than Brown's test [30, p. 100]. Recall that, in this example, Wilcoxon's test became sensitive to bias at $\Gamma = 1.7$. Despite correcting for testing twice, the adaptive test reported greater insensitivity to unmeasured bias than did Wilcoxon's test.

19.3.3 Example: Adaptive Inference for Lead in the Blood of Smokers

Returning to Fig. 19.1, there are many ways to test the null hypothesis of no treatment effect. We could analyze the differences in lead levels or the differences in the logs of the lead levels. We could ignore the number of cigarettes smoked by the smoker, or we could include that dose of smoking in the test statistic [45], as was done using T_{dose} in Sect. 18.4. We could use Wilcoxon's statistic, or Brown's [5] statistic, or the U-statistic $(m, \underline{m}, \overline{m}) = (8, 7, 8)$ that performs well with short-tailed errors, or the U-statistic $(m, \underline{m}, \overline{m}) = (8, 6, 7)$ that performs well with long-tailed errors. Each choice of statistic could lead to a different design sensitivity, and it is not obvious which statistic to choose.

The analysis in [31, Table 3] applied a total of twelve test statistics to Fig. 19.1, some with differences and others with differences of logs, some with doses others without, using two U-statistics, Wilcoxon's statistic and Brown's statistic. The analysis corrected for use of twelve statistics in parallel with (19.1), but now sharing α equitably among twelve statistics rather than two. Because this sharing of α used the joint distribution of twelve highly correlated statistics, it is nothing like using the Bonferroni inequality and allocating $\alpha/12$ to each statistic. At $\Gamma = 3.4$, the upper bound on the P-value from the adaptive test is 0.033 and it is produced by a statistic

considered in 19.2, namely $(m, \underline{m}, \overline{m}) = (8, 7, 8)$ applied to the differences in logs of lead levels, ignoring doses. Had we known in advance of seeing the data that this test statistic would be best, then we would have reported a bound on the P-value of 0.009 at $\Gamma = 3.4$, as in Sect. 19.2, so the price for looking at twelve tests rather than one is the move from 0.009 to 0.033. In contrast, $12 \times 0.009 = 0.11$ for correction using the Bonferroni inequality.

At $\Gamma = 3.4$, only three of the 12 individual, uncorrected analyses have a P-value bound less than or equal to 0.05. These three statistics all: (1) ignore doses, (2) use differences of logs of lead levels, and (3) avoid the various forms of Wilcoxon's statistic. There are two honest analyses: pick one statistic in advance, without looking at the data, or do an adaptive test that corrects for use of multiple test statistics. In this example, the correction for multiple testing is not large, and the task of guessing the best statistic is formidable.

19.4 Design Sensitivity and Dose–Response

19.4.1 Dose–Response and Evidence of Causality

Larger treatment effects are anticipated, perhaps correctly, at higher doses of treatment. Recall the debate quoted in Sect. 5.2.5. There, Austin Bradford Hill said an observed dose–response relationship is a consideration that may strengthen the claim that the treatment caused its ostensible effects. Kenneth Rothman disputed this. Rothman said bias may produce a dose–response relationship, whereas uniformly high doses may cause substantial effects with no dose–response relationship. Is a dose–response relationship merely one more association? Or is it evidence of causality?

In Sect. 5.2.5, this question was rephrased as an empirical question, not an abstract question, but rather a question about data at hand. Did an observed dose–response relationship lead to greater insensitivity to unmeasured bias, as measured by Γ? Table 5.5 focused just on the 12 high-dose matched pairs, finding greater insensitivity to unmeasured biases for these pairs than for all 22 matched pairs. So, in that specific sense, in that specific study, the dose–response relationship did strengthen the evidence that substantial biases would need to be present to explain away the ostensible treatment effect.

Is what we saw in Table 5.5 something we should expect to see? In Sect. 18.4, an answer was provided in terms of design sensitivity. In simple models, larger effects tend to be insensitive to larger biases. In simple models, if higher doses do produce larger effects—a big if—then a focus on treated individuals receiving higher doses will increase insensitivity to unmeasured biases in sufficiently large samples.

Suppose that we are considering a relationship between dose and response. What test statistics should we use? How does the choice of test statistic relate to the design

sensitivity? Perhaps the answer depends upon the sampling situation. If it does, can we hedge our bets using an adaptive test?

19.4.2 Example: Smoking and Periodontal Disease

Using data from the third National Health and Nutrition Examination Survey, NHANES III, Scott Tomar and Samira Asma [44] investigated cigarette smoking as a possible cause of periodontal disease. Figure 19.2 examines the same question using data from the more recent 2011–2012 NHANES.

Figure 19.2 depicts 441 daily cigarette smokers (S) matched to 441 nonsmoking controls (C). Smokers smoked every day for the previous 30 days, and 90% of them have smoked for 14.9 years or more. Nonsmokers smoked fewer than 100 cigarettes in their lives, do not smoke now, and had no tobacco use of any kind in the previous 5 days. The pairs were matched for age, gender, education in five categories, income, and black race. For measures of covariate balance produced by the match, see [37, Table 1 and Figure 1].

Following [44], periodontal disease is assessed at six locations on each of 28 teeth, excluding wisdom teeth if present, and a location exhibits periodontal disease if there is either a loss of attachment of $\geq$ 4 mm or a pocket depth of $\geq$ 4 mm. As in [38], the outcome is the percent of measurements indicating periodontal disease, 0–100, so matched pair differences range from -100 to 100. In Fig. 19.2, smokers

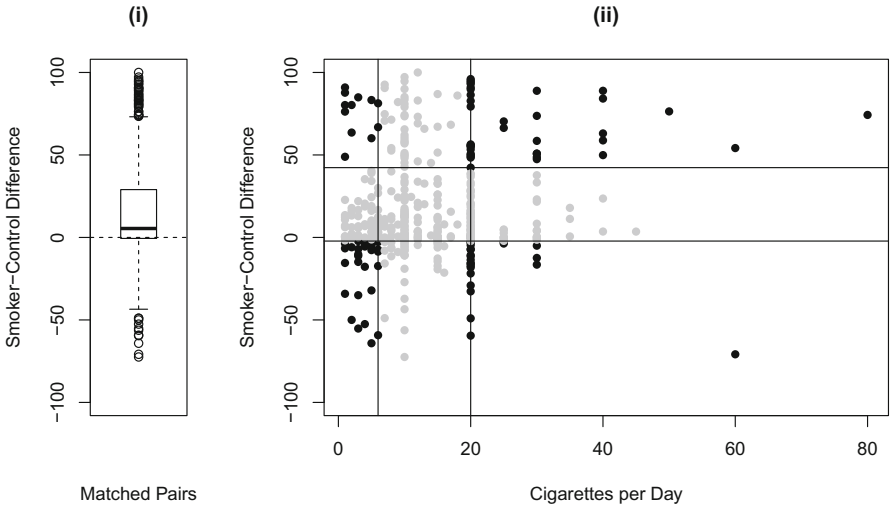

Fig. 19.2 Periodontal disease in 441 daily-smokers (S) and 441 matched nonsmoking controls (C). Panel (i) depicts the smoker-minus-control = S−C pair differences, Y_i. Panel (ii) plots these pair differences, Y_i, against the number d_i of cigarettes smoked per day by the smoker, deemphasizing in gray points that are strictly between the upper and lower quartiles

have more extensive periodontal disease than controls: the pair differences in panel (i) of Fig. 19.2 tend to be positive.

Panel (ii) of Fig. 19.2 plots the pair difference in periodontal disease against the number of cigarettes smoked per day by the smoker in each matched pair. The horizontal and vertical lines are at the upper and lower quartiles. Points in the outer corners are in black, the rest are in gray. In Panel (ii), larger pair differences tend to occur in pairs in which the smoker smoked heavily.

19.4.3 Ignoring Doses: Pair Differences in Periodontal Disease

Consider the $I = 441$ smoker-minus-control pair differences, Y_i, in periodontal disease in panel (i) of Fig. 19.2, ignoring the doses in panel (ii). When Wilcoxon's signed rank statistic is applied to the Y_i, a randomization test rejects the hypothesis H_0 of no effect with a vanishingly small P-value, and this test becomes sensitive to unmeasured bias at about $\Gamma = 2.8$ with a bound on the one-sided P-value of 0.061. If the U-statistic $(8, 7, 8)$ is used instead of Wilcoxon's statistic, the bound on the P-value is 0.00012 at $\Gamma = 2.8$ and is 0.049 at $\Gamma = 4.35$. So, with a suitable test statistic, rejection of H_0 at the 0.05 level in panel (i) of Fig. 19.2 is insensitive to fairly large unmeasured biases.

Do the doses, d_i, in panel (ii) of Fig. 19.2 provide any additional information?

19.4.4 The Crosscut Test

The crosscut test [36] is based on the four counts of black points in the outer corners of panel (ii) of Fig. 19.2; see Table 19.1. It is a test for dose–response, or more precisely for large effects at large doses. The test was designed for superior design sensitivity.

More generally, the crosscut test is based on the counts in the corners beyond the η quantiles of the dose d_i and the outcome Y_i variables, where $\eta = 0.25$ in Fig. 19.2. Cutting at the median, $\eta = 0.5$, yields the corner test of Paul Olmstead and John Tukey [24].

Table 19.1 Counts in the quartile corners for the crosscut test in Fig. 19.2, where Y_i is the pair difference in periodontal disease and d_i is cigarettes smoked per day by the smoker in a pair

Outcome	Cigarettes per day, d_i		
Y_i	$d_i \leq 7$	$d_i \geq 20$	Total
$Y_i \geq 29.00$	21	43	64
$Y_i \leq -0.60$	39	30	69
Total	69	73	133

In Table 19.1, we fix or condition upon the identities of the smokers in the matched pairs, and we ask if larger smoker-minus-control differences in periodontal disease are found in pairs in which the smoker smoked more cigarettes. If doses d_i were randomly assigned to pairs, and if Fisher's hypothesis H_0 of no effect of changing doses were true, then the counts inside Table 19.1 conditional on the marginal total counts would follow a hypergeometric distribution. Suppose instead that doses were not randomly assigned, but rather assigned in a biased manner such that pairs might differ in their odds of one dose rather than another by at most a factor of $\Gamma \geq 1$. Then, under H_0, the distribution of the counts in Table 19.1 are bounded by two extended hypergeometric distributions with parameters Γ and Γ^{-1}; see [36, Proposition 2] for specifics and proof. This leads to a test of H_0 and a sensitivity analysis for that test.

In Table 19.1, the randomization test, $\Gamma = 1$, of the hypothesis H_0 of no dose–response relationship yields a one-sided P-value of 0.0049, which could be doubled for a two-sided test. In terms of the extent of periodontal disease, heavy smokers differ more from their matched controls than do light smokers. The upper bound on this one-sided P-value is 0.049 if the bias in assigning doses to pairs is at most $\Gamma = 1.38$.

Figure 19.2 and Table 19.1 cut a continuous distribution at its quartiles. However, the crosscut test may be applied to the ordered categories of a contingency table; e.g., [36, Table 1].

Figure 19.2 and Table 19.1 cut at the quartiles, $\eta = 1/4$. Had we cut at the median instead, $\eta = 1/2$, then Table 19.1 would have included all $I = 441$ smoker-control pairs, but the results would be sensitive to smaller biases. For the median, $\eta = 1/2$, at $\Gamma = 1.38$, the upper bound on the one-sided P-value is 0.853, much larger than 0.049 for $\eta = 1/4$. Had we cut at 20% rather than 25%, $\eta = 1/5$ rather than $\eta = 1/4$, the results are insensitive to larger biases: at $\Gamma = 1.38$, the upper bound on the one-sided P-value is 0.0196, and at $\Gamma = 1.63$ it is 0.0487. What is the relationship between the cut, η, and sensitivity to unmeasured bias?

19.4.5 Design Sensitivity of the Crosscut Test

If there were a dose–response relationship produced by an actual treatment effect, and if there were no bias in assigning doses to pairs, then ask: How would the sensitivity analysis for the crosscut test turn out? As the sample size increases, $I \to \infty$, what is the limiting sensitivity to unmeasured bias? What is the design sensitivity $\widetilde{\Gamma}$? As expected, the answer depends upon the strength of the relationship between dose d_i and response Y_i. It also depends upon η, the quantile that defines the corners.

Table 19.2 gives the design sensitivity $\widetilde{\Gamma}$ for bivariate Normal (d_i, Y_i) with correlation ρ for several values of η. For more extensive tables of design sensitivities and simulated power of a sensitivity analysis for finite samples, including long-tailed distributions, see [36].

Table 19.2 Design sensitivity $\widetilde{\Gamma}$ of the crosscut test for bivariate Normal dose and response, (d_i, Y_i), with correlation ρ cutting at the outer η quantile

	Correlation ρ		
η	0.1	0.3	0.5
1/2	1.3	2.2	4.0
1/4	1.9	7.7	44.5
1/5	2.2	12.0	106.7
1/8	3.0	32.1	740.4

As expected, the design sensitivity in Table 19.2 increases as the correlation ρ between dose and response increases. In Table 19.2, the design sensitivity increases as η decreases. At correlation $\rho = 0.3$, the design sensitivity increases from $\widetilde{\Gamma} = 2.2$ when cutting at the median, $\eta = 1/2$, to $\widetilde{\Gamma} = 32.1$ when cutting at the outer eighths, $\eta = 1/8$.

As seen in [36, Table 3], the pattern in Table 19.2 also occurs in the long-tailed bivariate t-distribution with 3 degrees of freedom; however, the increase in design sensitivity as η decreases is somewhat less dramatic. As η decreases, fewer observations are in the outer corners, so fewer observations contribute to the crosscut test; however, in simulated samples with $I = 300$ or $I = 500$ pairs, the power of a sensitivity analysis remains high with $\eta = 1/5$; see [36, Table 4].

In brief, modest correlations, such as $\rho = 0.3$, yield large design sensitivities, $\widetilde{\Gamma}$, providing η is not large. Design sensitivity refers to the limit as the samples size increases, $I \to \infty$. Of course, for finite I, if η is too small, there will be very little data in the table analogous to Table 19.1. So, there is a trade-off that may be explored using adaptive inference, as discussed below.

19.4.6 The Stratified Crosscut Test

The analysis in Table 19.1 accounts for observed covariates by comparing smokers and matched controls who are similar in terms of observed covariates. Nonetheless, distinct pairs in Table 19.1 typically differ in terms of observed covariates: in one pair, smoker and control are both women aged 30, but in another pair, smoker and control are both men aged 70. Table 19.1 asks whether the doses predict the difference in outcomes in pairs matched for observed covariates.

A crosscut analysis can replace the one 2×2 table in Table 19.1 by a $2 \times 2 \times S$ table that forms S strata from some of the observed covariates already controlled by matching. Each matched pair is a count in this $2 \times 2 \times S$ table, but the pair of women aged 30 is now separated from the pair of men aged 70. With strata, the adjustment for observed covariates is done twice, in two different ways. In this case, for each $\Gamma \geq 1$, each of the S subtables is bounded by two extended hypergeometric distributions. The total of the S counts in the upper left corner cells of these S tables is the basis for the Mantel–Haenszel–Birch statistic [4, 18], and in

a sensitivity analysis, its null distribution is bounded by the convolution of extended hypergeometric distributions [36].

A crosscut analysis need not be applied to matched pairs. Instead, it might be applied to individuals who smoked in varying amounts. When applied to individuals rather than pairs, stratification becomes the sole adjustment for observed covariates.

19.4.7 An Adaptive Crosscut Test

Table 19.2 shows that sensitivity to unmeasured bias can vary dramatically with the choice of quantile η that defines the corners. The adaptive crosscut test [40, §6] cuts either the distribution of doses, d_i, or the distribution of outcomes, Y_i, or both distributions at four quantiles, $0 < \eta_1 < \eta_2 < 1/2 < 1 - \eta_2 < 1 - \eta_1 < 1$, rather than two quantiles. It then uses an adaptive version [40] of the Mantel–Haenszel–Birch statistic [4, 18] to consider several nonoverlapping 2×2 "corners" of the same joint distribution of dose and response. The adaptive test does two tests, one using just the "outer corners," the other using both "outer corners" and "inner corners," giving more weight to the "outer corners."

Motivated by the design sensitivities in Table 19.2, the adaptive test performs one test using a small η_1. Worried that a small η_1 entails using only a small part of the data, the adaptive test performs a second test with a larger η_2, hence a test that uses additional matched pairs. At each $\Gamma \geq 1$, the two tests equitably share the chance of a false rejection of a true null hypothesis of no effect in the sense of (19.1) and (19.2).

The adaptive test corrects for testing twice using the exact joint distribution of the two test statistics, but the correction is small because the two tests use the same data and are highly correlated. Despite testing twice, a level α adaptive test performed at Γ will falsely reject a true null hypothesis of no effect with probability at most α provided the bias in treatment assignment is at most Γ.

Suppose the doses d_i are cut at η_1 and η_2, but responses Y_i are cut only at η_1, resulting in a 3×5 table of counts of outcomes Y_i as rows by doses d_i as columns. This 3×5 table has a 2×2 subtable of outer corners at positions (1, 1), (1, 5), (3, 1), (3, 5). The 3×5 table also has a 2×2 subtable of inner corners at positions (1, 2), (1, 4), (3, 2), (3, 4). The two 2×2 tables do not overlap. These two 2×2 subtables are formed into a $2 \times 2 \times 2$ table, with two strata, outer corners or inner corners, and the adaptive Mantel–Haenszel–Birch test is applied to this $2 \times 2 \times 2$ table. The situation in which d_i is cut once but Y_i is cut twice is analogous, resulting in a 5×3 table, and a different $2 \times 2 \times 2$ table. A tabular example involving vitamin D and breast cancer is analyzed in [40, §6].

If the doses d_i and responses Y_i are both cut at η_1 and η_2, the result is a 5×5 table of counts. The 5×5 table has one 2×2 subtable of outer corners at positions (1, 1), (1, 5), (5, 1), (5, 5). The 5×5 table has two 2×2 subtables of inner corners, one at positions (1, 2), (1, 4), (5, 2), (5, 4), the other at positions (2, 1), (2, 5), (4, 1),

(4, 5). The three 2×2 subtables do not overlap, and they are combined into one $2 \times 2 \times 3$ table, to which the adaptive Mantel–Haenszel–Birch test is applied.

Consider cutting Fig. 19.2 not once at the quartiles, but twice at $\eta_1 = 1/5 = 0.2$ and $\eta_2 = 2/5 = 0.4$. Table 19.3 shows the counts for this double cut. The outer cuts in Table 19.3 are further out than in Table 19.1, excluding more pairs, while the inners cuts in Table 19.3 are further in than in Table 19.1, including more pairs. Can we let the data decide between the outer and inner cuts?

Table 19.4 selectively rearranges Table 19.3 into a $2 \times 2 \times 3$ table consisting of the outer corners in Table 19.3 plus two inner corners. Notice that the first of the three 2×2 tables in Table 19.4 has smaller counts than Table 19.1 but a larger odds ratio, $3.65 = (36 \times 28)/(12 \times 23)$ rather than $2.66 = (43 \times 39)/(21 \times 30)$. The cut at 1/5 yields a result that is less sensitive to bias than the cut at 1/4 in Table 19.1, consistent with the theoretical calculation in Table 19.2. Specifically, the upper bound on the one-sided P-value testing the hypothesis H_0 of no effect in Table 19.1 is 0.049 at $\Gamma = 1.38$, but the upper bound is 0.038 at $\Gamma = 1.55$ using just the first 2×2 table in Table 19.4. Of course, it could have been otherwise, and we could not know it would turn out this way before looking at the data.

Table 19.3 Counts in the corners for the crosscut test in Fig. 19.2, where d_i is cigarettes smoked per day by the smoker in a pair, and Y_i is the pair difference in periodontal disease

Outcome	Dose d_i, cigarettes per day					
Y_i	(0, 6]	(6, 10]	(10, 15)	[15, 20)	[20, ∞)	Total
[42.3, ∞)	12	29	7	5	36	89
[12.0, 42.3)	23	22	7	9	27	88
(0.9, 12.0)	24	23	6	9	25	87
(−2.2, 0.9]	22	25	3	15	23	88
($-\infty$, −2.2]	28	27	1	10	23	89
Total	109	126	24	48	134	441

Table 19.4 Conversion of the 5×5 counts in Table 19.3 into a $2 \times 2 \times 3$ table of outer and inner corners, cut at $\eta_1 = 1/5$ and $\eta_2 = 2/5$. In the adaptive test, the first 2×2 table is either used alone, or it is given double weight in combination with the other two 2×2 tables

Outcome	Dose d_i	
Outer Y_i, Outer d_i		
Y_i	(0, 6]	[20, ∞)
[42.3, ∞)	12	36
($-\infty$, −2.2]	28	23
Outer Y_i, Inner d_i		
Y_i	(6, 10]	[15, 20)
[42.3, ∞)	29	5
($-\infty$, −2.2]	27	10
Inner Y_i, Outer d_i		
Y_i	(0, 6]	[20, ∞)
[12.0, 42.3)	23	27
(−2.2, 0.9]	22	23

The adaptive crosscut test applied to Table 19.4 does two tests: (i) the test just done using the first 2×2 table in Table 19.4, and (ii) a Mantel–Haenszel–Birch test applied to the $2 \times 2 \times 3$ table, giving double weight to the first 2×2 table in Table 19.4. At $\Gamma = 1.55$, the adaptive test rejects H_0 at the 0.05-level, and it does so on the basis of test (i), not test (ii). The P-value of the adaptive test is 0.0499 at $\Gamma = 1.55$, rather than 0.038 at $\Gamma = 1.55$ if we knew to use the first 2×2 table alone; so, the price paid for an honest, adaptive test is not high.

19.4.8 The Crosscut Test and Evidence Factors

There is an intuitive sense in which the boxplot of pair differences Y_i in periodontal disease in panel (i) of Fig. 19.2 tells you little or nothing about the dose–response relationship between Y_i and the amount smoked, d_i, in panel (ii) of Fig. 19.2. Chapter 20 will explore this intuition and its implications in terms of the concept of evidence factors. More precisely, Chap. 20 will ask about the relationship between the dose–response test based on Table 19.1 and a smoker-minus-control comparison based on the pair differences Y_i alone, ignoring doses, say using Wilcoxon's signed rank test. It turns out that these two tests provide entirely separate, almost statistically independent pieces of information subject to different unobserved biases: it is as if two unrelated studies were conducted by different investigators, using different data, in different study designs.

19.5 Bahadur Efficiency of Sensitivity Analyses

19.5.1 Where Do Things Stand?

Throughout this chapter, the Pitman efficiency of a randomization test, $\Gamma = 1$, and the design sensitivity, $\widetilde{\Gamma}$, have favored different test statistics. Suppose, for instance, that the Y_i are independent observations from a Normal distribution with expectation τ and variance 1. Compare Noether's statistic [23] with $\eta = 1/3$ to Wilcoxon's statistic. In terms of Pitman efficiency, Noether's test is much worse, with relative efficiency 0.78 compared to Wilcoxon's test [30, Table 1], so Wilcoxon's test is a much better choice in a large randomized experiment studying a very small treatment effect. However, if $\tau = 1/2$, then the design sensitivity of Wilcoxon's test is $\widetilde{\Gamma} = 3.17$, much lower than Noether's statistic with $\widetilde{\Gamma} = 4.97$, so Noether's test is a much better choice in a large observational study with a meaningful treatment effect. This contrast is not surprising. The Pitman efficiency of a randomization test assumes there is no unmeasured bias and worries about efficient use of sample size, I. The design sensitivity $\widetilde{\Gamma}$ worries about unmeasured biases assuming the sample size is abundant, $I \to \infty$. Is there a way to simultaneously take account of bias and

sample size? Instead of jumping from no bias, $\Gamma = 1$, to the limiting bias, $\Gamma = \widetilde{\Gamma}$, can we discuss the efficiency of a sensitivity analysis performed at an intermediate Γ, with $1 < \Gamma < \widetilde{\Gamma}$?

19.5.2 Several Types of Efficiency

Suppose that we have two tests, T_1 and T_2, of the null hypothesis of no treatment effect, H_0. Which test should we use?

Chapter 15 computed the power of a sensitivity analysis, and of course T_1 and T_2 may have different powers in the same sampling situation. Recall the plots of the power of a sensitivity analysis in Figs. 15.2 and 15.3.

To compute the power, several things had to be specified: (i) the sampling situation, (ii) the test statistic, here either T_1 or T_2, (iii) the sample size I, (iv) the level of the test, conventionally level $\alpha = 0.05$, and finally (v) a value of the sensitivity parameter, Γ. Given (i)–(v), we apply a formula that returns the power, say Π. Here, Π is the probability that a sensitivity analysis performed at Γ will return an upper bound on the P-value of at most α. That is, Π is the probability that we will be able to report that rejection of H_0 at level α is insensitive to a bias in treatment assignment of Γ.

In earlier chapters, various "favorable" sampling situations were considered, that is, situations with a treatment effect and no unmeasured bias. Favorable situations were the focus because in a favorable situation we hope to report that conclusions are insensitive to small and moderate biases, measured by Γ. For instance, one possible favorable situation has Y_i independently drawn from a Normal distribution with expectation or treatment effect τ and variance 1. It is not the case that we know the truth or correct values in (i)–(v); rather, the power Π is a function of (i)–(v), so we gain insight into the power by varying (i)–(v) and seeing how the power changes.

We can read Figs. 15.2 and 15.3 backwards, specifying (i), (ii), (iv), (v), and the power Π, and finding the corresponding sample size, I. For instance, we might demand power of $\Pi = 0.8$. In words, we may ask: what sample size I gives us the power, $\Pi = 0.8$, that we demanded. Suppose we do this twice, once for T_1, once for T_2, obtaining two sample sizes, I_1 and I_2. So far as the sensitivity analysis at Γ is concerned, testing H_0 using T_1 with sample size I_1 is about the same as testing H_0 using T_2 with sample size I_2. In both cases, we have the same power Π, the same level α, at the same Γ in the same sampling situation. So we have created a metric, a currency, comparing T_1 and T_2, namely sample size. The better test gets the same work done with a smaller sample size. If $I_1 < I_2$, then T_1 is more efficient than T_2. If the relative efficiency of Wilcoxon's statistic to Noether's statistic were 1.28, or equivalently the relative efficiency of Noether's statistic to Wilcoxon's statistic were $0.78 = 1/1.28$, then Wilcoxon's statistic could get the same job done equally well with 78% of the sample size needed by Noether's statistic.

The quantity I_2/I_1 is the relative efficiency. If we had an infinite capacity and tolerance for such calculations, we would compute and memorize every relative efficiency for every setting of (i), (ii), (iv), (v), and Π, and we would know everything about the relative performance of T_1 and T_2. Lacking infinite capacity and tolerance for such calculations, we seek simplifications. Often, we find that the efficiency I_2/I_1 tends to a limit in one sense or another, and often the limit is easier to understand and remember than the infinite set of efficiencies I_2/I_1.

Speaking informally, Pitman efficiency considers an infinite sequence of treatment effects, $\tau_1, \tau_2, \ldots$, declining in magnitude toward zero effect, computes the efficiency I_2/I_1 for each τ_ℓ in this sequence, and if the sequence of efficiencies I_2/I_1 converges to a limit, then that limit is the Pitman asymptotic relative efficiency of T_1 and T_2.[4] In this calculation, Pitman efficiency holds the level, α, and the power, Π, fixed as we move through the decreasing sequence of τ_ℓ's, so both sample sizes, I_2 and I_1, are increasing to offset the declining treatment effect, but their ratio I_2/I_1 may be tending to a limit. Often, the limit of I_2/I_1 is the same for all levels α and powers Π, so Pitman efficiency produces a concise summary of all of the efficiencies, I_2/I_1. Pitman efficiency is so widely used that the terms "efficiency," "large-sample efficiency," and "asymptotic relative efficiency" are understood as references to Pitman efficiency.

Taken literally, the narrative of Pitman efficiency does not quite work for sensitivity analyses. As the size of the treatment effect becomes smaller, as the τ_ℓ approach zero, the study becomes sensitive to smaller and smaller biases, measured by Γ. Taken literally, the narrative of Pitman efficiency says that the smallest of biases is a giant problem, because I can no longer draw insensitive inferences about treatment effects of infinitesimally small size. In truth, no one cares about treatment effects of infinitesimal size. Treatment effects that are infinitesimally small are too small to care about; if you missed them all, then you have not missed much. In observational studies, it is a challenge to reach correct conclusions about treatment effects of moderate size, and those effects may be large enough to matter to someone.

Raghu Raj Bahadur [2] proposed a different form of asymptotic efficiency. He took a limit of I_2/I_1, but a different one. Unlike Pitman efficiency, Bahadur efficiency fixes the treatment effect, τ. For instance, we might settle on a treatment effect, τ, of moderate size, an effect that is large enough to matter to someone, an effect that has some hope of being insensitive to small or moderate biases measured by Γ. Also, Bahadur fixed the power, say $\Pi = 0.8$. Then, Bahadur considers a sequence of levels, $\alpha_1, \alpha_2, \ldots$, declining in magnitude toward zero, computes the efficiency I_2/I_1 for each α_ℓ in this sequence, and if the sequence of efficiencies I_2/I_1 converges to a limit, then that limit is the Bahadur relative efficiency of T_1

[4] Somewhat more precisely, the limit must be the same for every sequence of τ_ℓ tending to zero.

and T_2.[5] Typically, the Bahadur relative efficiency does not depend on the power, Π, but it does depend on the treatment effect τ.

Very small levels, α_ℓ, are practical in a way that very small treatment effects, τ_ℓ, are not. Imagine an investigator who plans to test many null hypotheses, perhaps in genomics, and plans to correct for multiple testing using the Bonferroni inequality. Such an investigator needs to test each hypothesis at a small uncorrected level, α_ℓ, and so has reason to want α_ℓ to decline as quickly as possible as the sample size increases. Missing an infinitesimally small treatment effect is not a big problem, but missing a big effect because many hypotheses were tested is a problem.

19.5.3 Using Bahadur Efficiency in Sensitivity Analyses

More to the point, a treatment effect τ of moderate size has some hope of being insensitive to small and moderate biases in an observational study. Unlike the narrative of Pitman efficiency, the narrative of Bahadur efficiency works for sensitivity analyses. The Bahadur efficiency of a randomization test, $\Gamma = 1$, is a familiar concept, so consider the Bahadur efficiency of a sensitivity analysis for $\Gamma > 1$. Suppose that the design sensitivity $\widetilde{\Gamma}_2$ of T_2 is larger than the design sensitivity $\widetilde{\Gamma}_1$ of T_1, $\widetilde{\Gamma}_2 > \widetilde{\Gamma}_1$. As Γ increases toward $\widetilde{\Gamma}_1$, the Bahadur relative efficiency of T_1 compared to T_2 will drop to zero. So, Bahadur efficiency will provide advice for $1 \leq \Gamma \leq \widetilde{\Gamma}_1$ that is compatible with the advice we already have about the extreme cases, $\Gamma = 1$ and $\widetilde{\Gamma}_1 < \Gamma < \widetilde{\Gamma}_2$.

Let us compare, one last time, the performance of Wilcoxon's test, T_1, and Noether's test with $\eta = 1/3$, T_2, when the Y_i are independently drawn from a Normal distribution with expectation $\tau = 1/2$ and variance 1. In a randomization test, $\Gamma = 1$, Wilcoxon's test is 35% more efficient than Noether's test, with I_2/I_1 tending to a limit of 1.35, the Bahadur relative efficiency. In the limit at $\Gamma = 1$, Noether's test requires a sample that is 35% larger to have the same performance as Wilcoxon's test. The situation reverses at $\Gamma = 2$, even though both statistics have design sensitivities well above 2, namely $\widetilde{\Gamma}_1 = 3.17$ for Wilcoxon's test, and $\widetilde{\Gamma}_2 = 4.97$ for Noether's test. At $\Gamma = 2$, the Bahadur relative efficiency is 0.68, so Noether's test can achieve with 68% of the sample what Wilcoxon's test needs the entire sample to achieve. At $\Gamma = 3$, the Bahadur relative efficiency is 0.03, because $\Gamma = 3$ is close to the design sensitivity of Wilcoxon's test, $\widetilde{\Gamma}_1 = 3.17$. For $3.17 < \Gamma < 4.97$, the Bahadur efficiency is zero. For these results and many others, see [34, Table 2].

In practical terms, the poor performance of statistics with small design sensitivities, $\widetilde{\Gamma}$, is already evident for Γ well below $\widetilde{\Gamma}$. This issue is quantified by the Bahadur efficiency of a sensitivity analysis.

[5] As in the previous footnote, we need a bit more, namely the same limit for every sequence of α_ℓ's tending to zero.

A remarkable theorem due to Robert Berk and Douglas Jones [3] concerns the Bahadur efficiency of adaptive inference of the type in (19.1). The theorem says that the adaptive test has the same Bahadur efficiency as the best of its several component tests. That is remarkable. The best test depends upon the sampling situation, and in practice you do not know the sampling situation. Despite your ignorance about which test will be best, the adaptive test has the same large sample performance as its best component. The theorem is remarkable because the adaptive test must correct for multiple testing, but in large samples that correction costs you nothing. Stated informally, by definition the component test with the largest Bahadur efficiency has a P-value that races to zero at a faster rate than its competitors, so in sufficiently large samples its P-value is all but certain to be ahead in this race, so the correction for multiple testing is negligible. In simulations of the finite-sample power of a sensitivity analysis, the adaptive test often lags its best component by a small distance, but the adaptive test stays well ahead of the other components [30, Table 3], suggesting that the Berk–Jones theorem has traction in problems of practical size.

Although the brief discussion here has focused on additive treatment effects with Normal errors, there is nothing special about this case, and similar calculations may be done with nonadditive effects and other error distributions [34, Tables 2 and 3]. Ashkan Ertefaie and colleagues [7] use Bahadur efficiency to examine the performance of strengthened instruments. Bikram Karmaker and colleagues [16] apply Bahadur efficiency to examine sensitivity analyses for evidence factors.

19.6 Further Reading

This chapter is a brief survey of a large topic. To build intuition, Figure 3 of [28, §3] is a good place to begin further reading: it plots Albers, Bickel and van Zwet's [1] function, $\text{abz}(y)$, against y for the Normal, logistic, double-exponential, and Cauchy distributions, indicating that $|Y_i|$ near zero are often sensitive to a bias of Γ when $|Y_i|$ a little further from zero are not.

For an introduction to adaptive sensitivity analysis, see [31]. For properties of the crosscut test and its sensitivity analysis, see [36], and for an adaptive crosscut test, see [40, §6]. For Bahadur efficiency of a sensitivity analysis, see [34].

19.7 Software

The `R` package `DOS2` contains a function, `senU`, that implements the U-statistic in Sect. 19.2, including confidence intervals and point estimates. With appropriate settings, `senU` performs Stephenson's test and the test based on the U-statistic that is essentially equivalent to Wilcoxon's signed rank test. The function `senWilcox`

in the DOS2 package performs the original Wilcoxon test with confidence intervals and point estimates. Both senU and senWilcox perform sensitivity analyses.

Packages sensitivitymv, sensitivitymw, and sensitivitymult implement the m-statistics in Sect. 19.2, although the functions senm and senmCI in sensitivitymult may be easiest to use. See also [35].

An interactive shinyApp teaching sensitivity analysis with m-statistics is available at https://rosenbap.shinyapps.io/learnsenShiny/. This is a graphical user interface, with text explanations, that analyzes a matched observational study by running the sensitivitymult in the background. In particular, you can see how changing the inner trim parameter, ι, changes sensitivity to unmeasured bias.

The senstrat package in R performs sensitivity analyses for unmatched comparisons stratified for covariates [39]. The package can use rank statistics or m-statistics.

The crosscut function in the R package DOS2 performs the crosscut test and its sensitivity analysis.

The mh function in the sensitivity2x2xk package performs a sensitivity analysis for the Mantel–Haenszel–Birch test [4, 18] in a $2 \times 2 \times S$ table. The function mh may be used to perform a sensitivity analysis for the stratified crosscut test [36].

The crosscut test was illustrated using a cut at the outer quartiles, $\eta = 0.25$. The adaptive crosscut test in [40, §6] uses two or more cuts at once, say $\eta = 1/5$ and $\eta = 1/3$. This test may be performed by computing the relevant $2 \times 2 \times S$ table and applying to it the adaptmh function in the sensitivity2x2xk package. The examples for adaptmh use an example in [40] to perform an adaptive crosscut test.

19.8 Data

The blood lead data in Fig. 19.1 are contained in the R package DOS2 as smokerlead, and some of the calculations are reproduced in its help-file.

The periodontal disease and smoking data in Fig. 19.2 are contained in the R package DOS2 as periodontal and its analysis appears in the help-file for the crosscut function in that package.

Bibliography

1. Albers, W., Bickel, P.J., van Zwet, W.R.: Asymptotic expansions for the power of distribution free tests in the one-sample problem. Ann. Stat. **4**, 108–156 (1976)
2. Bahadur, R.R.: Rates of convergence of estimates and test statistics. Ann. Math. Stat. **38**, 303–324 (1967)
3. Berk, R.H., Jones, D.H.: Relatively optimal combinations of test statistics. Scand. J. Stat. **5**, 158–162 (1978)

4. Birch, M.W.: The detection of partial association, I: the 2 x 2 case. J. R. Stat. Soc. B. **26**, 313–324 (1964)
5. Brown, B.M.: Symmetric quantile averages and related estimators. Biometrika **68**, 235–242 (1981)
6. Conover, W.J., Salsburg, D.S.:Locally most powerful tests for detecting treatment effects when only a subset of patients can be expected to 'respond' to treatment. Biometrics **44**, 189–196 (1988)
7. Ertefaie, A., Small, D.S., Rosenbaum, P.R.: Quantitative evaluation of the trade-off of strengthened instruments and sample size in observational studies. J. Am. Stat. Assoc. **113**, 1122–1134 (2018)
8. Gastwirth, J.L.: On robust procedures. J. Am. Stat. Assoc. **61**, 929–948 (1966)
9. Groeneboom, P., Oosterhoff, J.: Bahadur efficiency and small-sample efficiency. Int. Stat. Rev. **49**, 127–141 (1981)
10. Groeneveld, R.A.: Asymptotically optimal group rank tests for location. J. Am. Stat. Assoc. **67**, 847–849 (1972)
11. Hansen, B.B., Rosenbaum, P.R., Small, D.S.: Clustered treatment assignments and sensitivity to unmeasured biases in observational studies. J. Am. Stat. Assoc. **109**, 133–144 (2014)
12. Heller, R., Rosenbaum, P.R., Small, D.: Split samples and design sensitivity in observational studies. J. Am. Stat. Assoc. **104**, 1090–1101 (2009)
13. Hoeffding, W.: A class of statistics with asymptotically Normal distribution. Ann. Math. Stat. **19**, 293–325 (1948)
14. Huber, P.J.: Robust estimation of a location parameter. Ann. Math. Stat. **35**, 73–101 (1964)
15. Huber, P.J.: Robust Statistics. Wiley, New York (1981)
16. Karmakar, B., French, B., Small, D.S.: Integrating the evidence from evidence factors in observational studies. Biometrika **106**, 353–367 (2019)
17. Lehmann, E.L.: Nonparametrics. Holden Day, San Francisco (1975). Reprinted Springer, New York (2006)
18. Mantel, N., Haenszel, W.: Statistical aspects of the analysis of data from retrospective studies of disease. J. Nat. Cancer Inst. **22**, 719–748 (1959)
19. Maritz, J.S.: A note on exact robust confidence intervals for location. Biometrika **66**, 163–166 (1979)
20. Maritz, J.S.: Distribution-Free Statistical Methods. Chapman and Hall, London (1995)
21. Markowski, E.P., Hettmansperger, T.P.: Inference based on simple rank step score statistics for the location model. J. Am. Stat. Assoc. **77**, 901–907 (1982)
22. Nikitin, I.: Asymptotic Efficiency of Nonparametric Tests. Cambridge, NewYork (1995)
23. Noether, G.E.: Some simple distribution-free confidence intervals for the center of a symmetric distribution. J. Am. Stat. Assoc. **68**, 716–719 (1973)
24. Olmstead, P.S., Tukey, J.W.: A corner test for association. Ann. Math. Stat. **18**, 495–513 (1947)
25. Rosenbaum, P.R.: Using quantile averages in matched observational studies. J. R. Stat. Soc. C (Appl. Stat.) **48**, 63–78 (1999)
26. Rosenbaum, P.R.: Sensitivity analysis for m-estimates, tests, and confidence intervals in matched observational studies. Biometrics **63**, 456–464 (2007)
27. Rosenbaum, P.R.: Confidence intervals for uncommon but dramatic responses to treatment. Biometrics **63**, 1164–1171 (2007)
28. Rosenbaum, P.R.: Design sensitivity and efficiency in observational studies. J. Am. Stat. Assoc. **105**, 692–702 (2010)
29. Rosenbaum, P.R.: A new U-statistic with superior design sensitivity in matched observational studies. Biometrics **67**, 1017–1027 (2011)
30. Rosenbaum, P.R.: An exact adaptive test with superior design sensitivity in an observational study of treatments for ovarian cancer. Ann. Appl. Stat. **6**, 83–105 (2012)
31. Rosenbaum, P.R.: Testing one hypothesis twice in observational studies. Biometrika **99**, 763–774 (2012)

32. Rosenbaum, P.R.: Impact of multiple matched controls on design sensitivity in observational studies. Biometrics **69**, 118–127 (2013)
33. Rosenbaum, P.R.: Weighted m-statistics with superior design sensitivity in matched observational studies with multiple controls. J. Am. Stat. Assoc. **109**, 1145–1158 (2014)
34. Rosenbaum, P.R.: Bahadur efficiency of sensitivity analyses in observational studies. J. Am. Stat. Assoc. **110**, 205–217 (2015)
35. Rosenbaum, P.R.: Two R packages for sensitivity analysis in observational studies. Observ. Stud. **1**, 1–17 (2015)
36. Rosenbaum, P.R.: The crosscut statistic and its sensitivity to bias in observational studies with ordered doses of treatment. Biometrics **72**, 175–183 (2016)
37. Rosenbaum, P.R.: Using Scheffe projections for multiple outcomes in an observational study of smoking and periodontal disease. Ann. Appl. Stat. **10**, 1447–1471 (2016)
38. Rosenbaum, P.R.: The general structure of evidence factors in observational studies. Stat. Sci. **32**, 514–530 (2017)
39. Rosenbaum, P.R.: Sensitivity analysis for stratified comparisons in an observational study of the effect of smoking on homocysteine levels. Ann. Appl. Stat. **12**, 2312–2334 (2018)
40. Rosenbaum, P.R., Small, D.S.:An adaptive Mantel-Haenszel test for sensitivity analysis in observational studies. Biometrics **73**, 422–430 (2017)
41. Serfling, R.J.: Approximation Theorems of Mathematical Statistics. Wiley, New York (1980)
42. Shaper, A.G., Pocock, S.J., Walker, M., Wale, C.J., Clayton, B., Delves, H.T., Hinks, L.: Effects of alcohol and smoking on blood lead in middle-aged British men. Brit. Med. J. **284**, 299–302 (1982)
43. Stephenson, W.R.: A general class of one-sample nonparametric test statistics based on subsamples. J. Am. Stat. Assoc. **76**, 960–966 (1981)
44. Tomar, S.L., Asma, S.: Smoking-attributable periodontitis in the United States: findings from NHANES III. J. Periodont. **71**, 743–751 (2000)
45. van Eeden, C.: An analogue, for signed rank statistics, of Jureckova's asymptotic linearity theorem for rank statistics. Ann. Math. Stat. **43**, 791–802 (1972)
46. Wilcoxon, F.: Individual comparisons by ranking methods. Biometrics **1**, 80–83 (1945)
47. Zhao, Q.:On sensitivity value of pair-matched observational studies. J. Am. Stat. Assoc. **114**, 713–722 (2019)
48. Zhao, Q., Small, D.S., Rosenbaum, P.R.: Cross-screening in observational studies that test many hypotheses. J. Am. Stat. Assoc. **113**, 1070–1084 (2018)
49. Zhao, Q., Small, D.S., Su, W.: Multiple testing when many p-values are uniformly conservative, with application to testing qualitative interaction in educational interventions. J. Am. Stat. Assoc. **114**, 1291–1304 (2019)

Part IV
Enhanced Design

Chapter 20
Evidence Factors

Abstract An observational study has two evidence factors if it permits two essentially independent tests of the null hypothesis of no treatment effect, where each test is unaffected by some unmeasured bias that would invalidate the other test. Because the two tests are essentially independent, the evidence they provide—their hypothesis tests and sensitivity analyses—may be combined using meta-analytic techniques, as if the two tests were performed in different studies by different investigators.

> [We should] proceed only from tangible premises which can be subjected to careful scrutiny, and . . . trust rather to the multitude and variety of . . . arguments than to the conclusiveness of any one. [Our] reasoning should not form a chain which is no stronger than its weakest link, but a cable whose fibers may be ever so slender, provided they are sufficiently numerous and intimately connected.
>
> Charles Sanders Peirce [24]

> The combined evidence on a question that has to be decided mainly from observational studies will usually consist of a heterogeneous collection of results of varying quality, each bearing on some consequence of the causal hypothesis . . . [The investigator] cannot avoid an attempt to weigh the evidence for and against, since some results are so vulnerable to bias that they should be given low weight even if supported by routine tests of significance.
>
> William G. Cochran [3, pp. 252–253]

20.1 What Are Evidence Factors?

20.1.1 A Replication Should Disrupt a Plausible Bias

Biases can replicate, and they often do. When replicating an observational study, we try to replicate the treatment effect without replicating whatever biases were also present in the initial study; see Sect. 4.5. Replicating a treatment effect without replicating biases entails studying the same effect independently, in a new context, in

© Springer Nature Switzerland AG 2020

P. R. Rosenbaum, *Design of Observational Studies*, Springer Series in Statistics,
https://doi.org/10.1007/978-3-030-46405-9_20

a new way [27, 39, 40]. The reappearance of a treatment effect when studied again is not convincing on its own; rather, we are convinced by its repeated reappearance in new contexts that have deliberately and thoughtfully disrupted some likely sources of bias. As noted in Sect. 4.5, Mervyn Susser [39, p. 148] wrote: "The strength of this approach rests on the fact that diverse approaches produce similar results."

Attempts to replicate treatment effects often fail. An effect apparently present in one context is apparently absent in another. Attempts to replicate effects while disrupting biases often fail, so the outcome of an attempt to replicate a finding is not a foregone conclusion. If the same effect reappears in a new study conducted in a new way, it is news.

20.1.2 Replication and Evidence Factors

Understanding replication in this way, now ask: Can one study replicate itself? Can one study provide two independent tests of the null hypothesis of no effect, such that the biases attending the first test might be quite different from the biases attending the second test? Because the two tests are statistically independent, despite using the same data, the finding of the second test is not a foregone conclusion given the outcome of the first test. Because the two tests are susceptible to different biases, the two tests will falsely concur only if the two biases conspire to produce the same false impression of a treatment effect. Can a single study contain its own disruptive replication?

Statistical independence of the two tests is important, but it is not the key element. If all we wanted were two statistically independent tests of one hypothesis of no treatment effect, then we could simply split the sample in half at random, testing independently in each half. There is no point in doing this. Two random halves of a single study are affected by the same biases—precisely nothing has been done to disrupt the biases. If one random half is biased, so is the other, and in the same way. In this context, random splitting is perfectly mistaken: it is the only way you can be absolutely certain that the two independent tests are biased in exactly the same way.

If a single study provides two statistically independent tests of one hypothesis of no treatment effect, and if these two tests are susceptible to different biases in treatment assignment, then there are two evidence factors [30, 31, 36]. Ideally, the sensitivity analyses for these two factors also provide nonoverlapping information, in a sense that will be developed in the current chapter.

20.1.3 Example: Smoking and Periodontal Disease

Consider again the example in Sect. 19.4 concerning smoking as a possible cause of periodontal disease. In particular, reexamine the two panels of Fig. 19.2. There was a strong intuition that the boxplot of the smoker-minus-control pair differences, Y_i,

in periodontal disease in panel (i) of Fig. 19.2 told you little or nothing about the relationship in panel (ii) of Y_i and the amount smoked by the smoker, d_i. There could have been a shift towards positive Y_i in panel (i) of Fig. 19.2, but no relationship between Y_i and d_i in panel (ii). Conversely, Y_i could tend to increase with d_i in panel (ii), yet Y_i could be symmetric about zero in panel (i). Are there two separate pieces of information in Fig. 19.2? If so, in what sense is this information separate? Can two separate pieces of information be combined to form strengthened evidence of an effect of smoking on periodontal disease? Can their sensitivity analyses be combined?

20.1.4 Issues Developed in This Chapter

If we had two evidence factors, life would be simple. We could analyze one study twice – use the same data twice – and act as if we had done two separate studies, replicating our own work. The conditions that permit this require some care, first in understanding the conditions, then in designing studies in which the conditions are satisfied. There are several issues developed in this chapter:

- Strictly independent evidence factors arise for certain comparisons in randomized experiments, providing certain rank statistics are applied to data free of ties. For instance, imagine a randomized experiment in Fig. 19.2 in which 441 untied doses, d_i, are randomly permuted and assigned to the $I = 441$ matched pairs, and then one person in each pair is picked at random to be treated with that dose, the other person receiving control. If the $I = 441$ absolute pair differences in outcomes, $|Y_i|$ are also untied and never zero, and if Fisher's hypothesis of no dose/treatment effect were true, then in this randomized experiment, a signed rank statistic and the crosscut statistic in Sect. 19.4 are independent [30].
- Strict independence of two factors is easily lost, perhaps due to ties in d_i, or ties in $|Y_i|$, or due to choosing the wrong pair of rank statistics, or due to replacing a signed rank statistic by a mean or an m-statistic. This seems unpleasant and tedious. Small changes in the set-up seem to invalidate the whole approach. However, it turns out that we do not need strict independence, and a weaker but equally useful result is available without worries about ties or rank statistics.
- What replaces strict independence of two factors? Each factor provides a P-value testing the null hypothesis H_0 of no treatment effect, say P_1 and P_2. Under H_0 in a randomized experiment, it suffices that this pair of two P-values be stochastically larger than the uniform distribution on the unit square [2, 31]. Although two P-values can be dependent and yet also stochastically larger than the uniform distribution on the square, the latter condition is what is needed in hypothesis testing – that is, we do not need strict independence. We need not worry about ties or rank statistics to obtain two P-values that are stochastically larger than the uniform distribution on the square; rather, it suffices to focus on certain symmetries in study design; see Sect. 20.4 and [36].

- We do not only want to combine P-values in randomized experiments, $\Gamma = 1$, but also to combine bounds on P-values from sensitivity analyses in observational studies, $\Gamma > 1$. What is required to do this? When there is bias from an unmeasured covariate $\mathbf{u}$, we cannot compute the true P-values, (P_1, P_2), because we do not have access to $\mathbf{u}$. Suppose the bias in treatment assignment is bounded in magnitude by the value of a (perhaps two-dimensional) sensitivity parameter, and given this supposition, that we have computed separate bounds on the two P-values, $\bar{P}_1 \geq P_1$ and $\bar{P}_2 \geq P_2$. If the unknown, true P-values, (P_1, P_2), were jointly stochastically larger than the uniform distribution on the unit square, then so are the known bounds, $(\bar{P}_1, \bar{P}_2)$, simply because they are deterministically larger than (P_1, P_2). To treat the bounds $(\bar{P}_1, \bar{P}_2)$ as valid, if conservative, joint P-values, we need to show that the true, unknown P-values, (P_1, P_2), are stochastically larger than the uniform distribution on the unit square. It would simplify life if separate sharp bounds on P_1 and P_2 were also sharp bounds on (P_1, P_2) jointly; then, there would be no loss in simply combining two separate sensitivity analyses.

Sections 20.2 and 20.3 motivate and clarify these issues in the context of two scientific examples. The slightly technical appendix to this chapter, Sect. 20.4, presents a simplified version of the general argument from [36].

20.2 The Simplest Nontrivial Case: Renyi's Partial Ranks

20.2.1 DNA Damage Among Tannery Workers

Workers who tan leather are often exposed to trivalent chromium, which may damage DNA. Zhang and colleagues [44] compared three groups: 30 male factory workers exposed to high levels of chromium by working in the factory's tannery (E1), 30 men possibly exposed to much lower levels of chromium by working in the same factory's finishing department (E2), and 30 male controls believed to have no substantial exposure to chromium (C). The three groups were of similar age, and in Table 20.1 they are grouped in 30 blocks of size 3 to control cigarette smoking as closely as possible.

In Table 20.1, DNA damage is measured by the mean tail moment of the comet assay, which is often used to measure DNA strand breaks [5, 23]. The comet assay places DNA in an electric field that pulls smaller, broken pieces of DNA away from intact DNA, producing an image that resembles the tail of a comet. Broken pieces are often pulled to the tail, intact DNA often remains in the comet's head, so the magnitude of the tail is taken as a measure of DNA damage. Higher values of the mean tail moment (mtm) indicate a larger tail and more extensive DNA damage (Fig. 20.1).

Table 20.1 offers several comparisons, and perhaps several comparisons are needed. Perhaps everyone in the factory is exposed to trivalent chromium to some

Table 20.1 DNA damage among two groups, E1 and E2, each containing 30 tannery workers, and 30 controls, in 30 blocks of three individuals. Values are the mean tail moment (mtm) of the comet assay. Source Zhang et al. [44]

	Tannery workers		
Department	Tannery	Finishing	Controls
Block	E1	E2	C
1	3.02	1.79	1.45
2	6.60	2.15	2.48
3	4.32	5.15	0.20
4	10.70	7.46	0.35
	⋮		
30	5.57	2.59	0.29

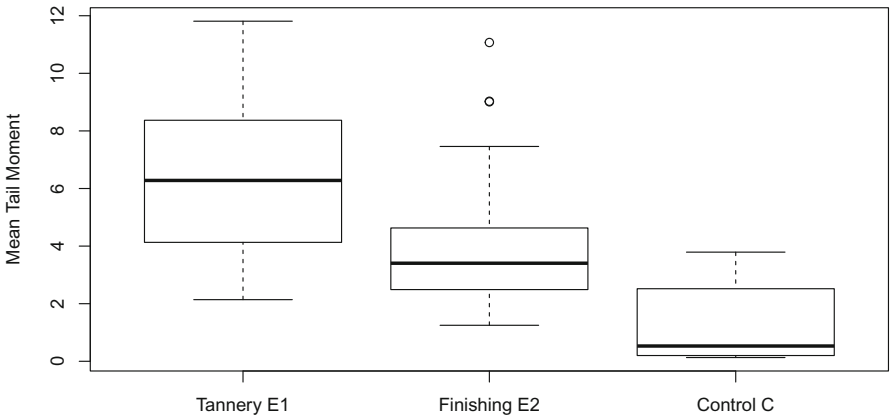

Fig. 20.1 Mean tail moment of the comet assay for 30 workers in the tannery department (E1), 30 workers in the finishing department (E2), and 30 unexposed controls (C). In the finishing department, exposure to trivalent chromium is much lower than in the tannery department

degree, so the comparison of groups E1 and E2 understates the harm done by high levels of chromium in the tannery (Fig. 20.2). Perhaps everyone in the factory is exposed to various toxins in various degrees, so a comparison with controls from other sources overstates the harm done by working in the factory's tannery. Are there two evidence factors in Table 20.1?

20.2.2 Renyi's Partial Ranks in a Randomized Experiment Under the Null Hypothesis

The simplest case is a randomized experiment in which Fisher's null hypothesis, H_0, of no treatment effect is assumed to be true for the purpose of testing it. In this simplest case, randomization permutes individuals but does not change their outcomes. In a blocked randomized experiment, randomization permutes individuals within blocks.

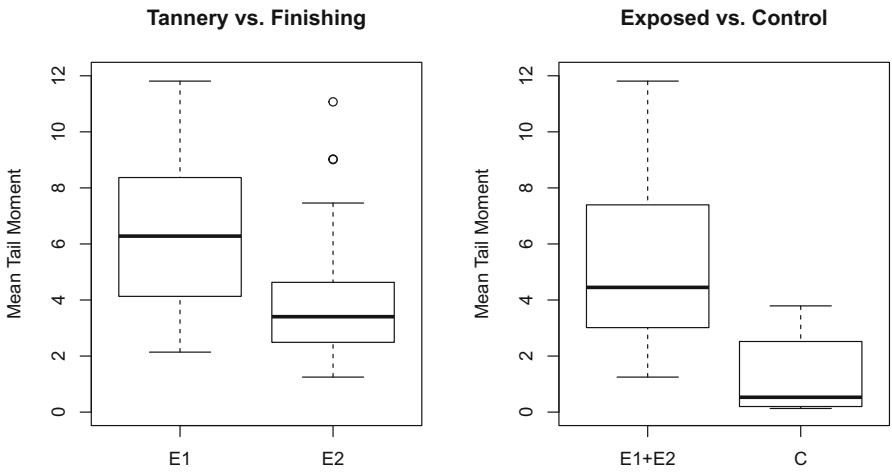

Fig. 20.2 Rearrangement of Fig. 20.1 into two comparisons, one comparing 30 workers in the tannery department (E1) to 30 workers in the finishing department (E2), the other comparing 60 workers to 30 unexposed controls (C)

Table 20.2 Partial ranks are independent in a randomized experiment in the absence of a treatment effect. The $6 = 3!$ permutations of the 3 people in the first block in Table 20.1. In a randomized experiment, each of the 6 permutations has probability 1/6, so the ranks in the last two columns are independent

	Outcomes			Partial ranks	
Permutation	E1	E2	C	E2 in E1,E2	C in E1,E2,C
1	3.02	1.79	1.45	1	1
2	1.79	3.02	1.45	2	1
3	3.02	1.45	1.79	1	2
4	1.45	3.02	1.79	2	2
5	1.79	1.45	3.02	1	3
6	1.45	1.79	3.02	2	3

If H_0 were true in block 1 of Table 20.1, and if randomization permuted individuals independently within each block, then for block 1 we would see one row of the six rows of Table 20.2, with each row having probability $1/6$. There are six rows in Table 20.2 because we can arrange 3 people in $3! = 3 \times 2 \times 1 = 6$ possible ways, and randomization picks one of these.

Define partial ranks as follows. In a block, let the partial rank of individual E2 be the rank of individual E2 among individuals {E1,E2}, so this rank is 1 or 2. In a block, let the partial rank of individual C be the rank of individual C among individuals {E1,E2,C}, so this rank is 1 or 2 or 3. Table 20.2 shows the partial ranks for the six equally probable random assignments of individuals to treatment positions.

Table 20.3 Joint distribution of the partial ranks in Table 20.2, under the null hypothesis H_0 of no effect in a randomized experiment

Rank of E2 in E1,E2	Rank of C in E1,E2,C			Total
	1	2	3	
1	1/6	1/6	1/6	1/2
2	1/6	1/6	1/6	1/2
Total	1/3	1/3	1/3	1

In Table 20.2, the partial ranks of E2 and C are independent; see Table 20.3. For instance, the conditional probability is 1/3 that the partial rank of C equals 3 given the partial rank of E2, whether the partial rank of E2 is 1 or 2. This is the simplest illustration of a more general result of Alfred Renyi; see, for instance, Alam [1] or Resnick [26, §4.3.1]. There are many useful generalizations and modifications of Renyi's result [7, 20, 25, 42].

Although Tables 20.2 and 20.3 are the simplest case, they provide one useful method for the analysis of Table 20.1.

20.2.3 Wilcoxon's Stratified Rank Sum Test

Frank Wilcoxon developed two randomization tests [41], the signed rank test for matched pairs and the rank sum test for comparison of two groups. The signed rank test was discussed in Sect. 2.3.3, Chaps. 3 and 15. The signed rank statistic in its usual form is not applicable to the matched triples in Table 20.1.

A completely randomized experiment begins with N individuals available for treatment, randomly selects m individuals for treatment 1, assigning the remaining $N - m$ individuals to treatment 2. Viewing the control as treatment 1 in Table 20.2, so $N = 3$ and $m = 1$, there are three choices of one individual in block 1 to receive the control, and each of these choices has probability $1/3$ in a completely randomized experiment. In general, there are $N!/\{m! \times (N-m)!\}$ ways to pick m individuals for treatment 1 from N individuals and, in a completely randomized experiment, each of the $N!/\{m! \times (N-m)!\}$ possible treatment assignments has the same probability, namely $[N!/\{m! \times (N-m)!\}]^{-1}$. In Table 20.2 with $N = 3$ and $m = 1$, there are $N!/\{m! \times (N-m)!\} = 3!/\{1! \times (3-1)!\} = 6/2 = 3$ possible choices of the one control.

If Fisher's sharp null hypothesis H_0 of no treatment effect were true in a completely randomized experiment, then each individual would exhibit the same response whether that individual was assigned to treatment group 1 or treatment group 2. As a result, when there is no treatment effect in a completely randomized experiment, randomization labels some individuals as treatment 1, others as treatment 2, but it does not alter their responses. In a completely randomized experiment with no treatment effect, the responses in treatment group 1 are a random sample

of m responses from among N responses.[1] If Table 20.2 described a completely randomized experiment with no treatment effect, the response of the $m = 1$ individual selected for control from among the $N = 3$ individuals would be one of the $N = 3$ numbers 3.02, 1.79, and 1.45, each having probability 1/3.

In a completely randomized experiment, Wilcoxon's rank sum statistic ranks the responses of all N individuals from 1 to N, using average ranks for ties, and the statistic is the sum of the m ranks for the m individuals in treatment group 1. Wilcoxon's rank sum test asks: Do the m ranks in treatment group 1 look like m numbers picked at random from the N ranks? Or are the ranks in treatment group 1 too high to be a random sample from the N ranks? Or too low?

In parallel with Sect. 2.3.3 for the signed rank statistic, randomization creates the distribution of Wilcoxon's rank sum statistic under Fisher's hypothesis H_0 of no effect. If Table 20.2 described a completely randomized experiment with no effect, then the rank of the response of the $m = 1$ individual selected for control from among the $N = 3$ individuals would be one of the $N = 3$ numbers 3, 2, and 1, each having probability 1/3, so that, in this simple case, the distribution of the rank sum statistic attaches probability 1/3 to 3, 1/3 to 2, and 1/3 to 1. Of course, this is the same as the distribution of the partial rank for the control in Table 20.3.

20.2.4 Two Evidence Factors Using Wilcoxon's Rank Sum Test

Can one of Wilcoxon's statistics be used in Table 20.1? Unlike a completely randomized experiment, Table 20.1 has 30 blocks, so Wilcoxon's rank sum statistic is not directly applicable [9]. Unlike a paired randomized experiment, Table 20.1 has three individuals in each of 30 blocks, not pairs, so Wilcoxon's signed rank statistic is not directly applicable.

A randomized experiment analogous to Table 20.1 is a randomized block experiment. In a randomized block experiment, each of the 30 blocks in Table 20.1 would be randomized independently, in essence repeating the process in Table 20.2 thirty times. In each block, one individual is picked at random for control, and then one of the remaining individuals is picked at random for group E2.

Each block yields a rank sum statistic comparing the control C to {E1,E2,C}, thirty statistics in all. Indeed, the first of the thirty rank sum statistics, the statistic for block 1, is the partial rank for the control group in Table 20.2. Each block yields another rank sum statistic comparing E2 to {E1,E2}. The first of the thirty rank sum statistics is the partial rank for the E2 in Table 20.2.

In Table 20.1, all thirty blocks have the same structure, one individual in each of three groups. When all blocks have the same structure, the design is said to be

[1]More precisely, in a completely randomized experiment with no treatment effect, there is a population of N fixed responses, and the m responses in treatment group 1 are a simple random sample of size m, drawn without replacement, from the population of size N.

a balanced design. In a balanced design, the stratified Wilcoxon rank sum statistic is simply the sum of the Wilcoxon rank sum statistics for the individual blocks [18, pp. 132–135].

In a randomized block experiment with no treatment effect, the stratified Wilcoxon statistic has a distribution created by randomization. Suppose, for example, that Table 20.1 was from a randomized block design with no treatment effect. Then the stratified rank sum statistic comparing C to {E1, E2, C} would be the sum of 30 independent rank sum statistics, each with $N = 3$ individuals with $m = 1$ selected for control. The expected value of each of the 30 rank sum statistics is $1 \times (1/3) + 2 \times (1/3) + 3 \times (1/3) = 2$, so the expectation of the sum of 30 such statistics is $30 \times 2 = 60$. The variance of each of the 30 rank sum statistics is $(1-2)^2 \times (1/3) + (2-2)^2 \times (1/3) + (3-2)^2 \times (1/3) = 2/3$, so the variance of the sum of 30 such independent statistics is $30 \times 2/3 = 20$, and the standard deviation is $\sqrt{20} = 4.472$. In fact, the stratified rank sum for controls is 35, not 60, with a standardized deviate of $(35-60)/\sqrt{20} = -5.59$ and an approximate P-value from a Normal distribution of 1.134×10^{-8}. The exact null distribution may be used instead of the Normal approximation. In a randomized experiment, we would judge the null hypothesis of no treatment effect to be implausible, with the 30 controls having less DNA damage than the 60 factory workers, the one-sided P-value being very small.

There is a second Wilcoxon rank sum statistic comparing E2 to {E1, E2}, or the sum of the partial ranks for E2. In each of 30 strata, the ranks are 1 or 2. In a stratified randomized experiment with no effect, the expectation of the stratified statistic is $30 \times \{1 \times (1/2) + 2 \times (1/2)\}$ or $30 \times 1.5 = 45$ and the variance is $30 \times \{.5^2 \times (1/2) + .5^2 \times (1/2)\}$ or $30 \times .25 = 7.5$. In fact, the stratified Wilcoxon statistic comparing E2 to {E1, E2} is 37, not 45, with standardized deviate $(37-45)/\sqrt{7.5} = -2.921$ with approximate one-sided P-value 0.00174. In a randomized experiment, we would judge the null hypothesis of no treatment effect to be implausible, with the 30 factory workers in the finishing room having less DNA damage than the 30 factory workers in the tannery, the one-sided P-value being very small.

So, two tests of the null hypothesis of no effect, H_0, would both provide fairly strong evidence against H_0 in a randomized experiment. The remarkable thing is that these two tests would be statistically independent in a randomized block experiment were H_0 true. This follows from Table 20.3 or equivalently from Renyi's result about partial ranks. The study has replicated itself, using two independent tests that are likely to be affected by different biases.

Because the two P-values are independent, they can be combined using various methods for combining independent P-values, such as Fisher's method. Combining P-values of 1.134×10^{-8} and 0.00174 using Fisher's method yields a much smaller combined P-value of 5.072×10^{-10}.

In brief, these calculations exhibit the simplest, nontrivial example of two evidence factors.

20.2.5 Limitations of Partial Ranks

So far, we have considered the simplest case: two evidence factors from three groups in randomized balanced blocks of size three, employing the stratified Wilcoxon statistic. There are straightforward extensions. Renyi's result is not restricted to two independent factors from three groups [1, 7, 20]. Wilcoxon's statistic sums the ranks, but the same argument applies if the statistic sums functions of the ranks. For instance, Chap. 17 considered transformed ranks proposed by Conover and Salsburg [6] that increase power and design sensitivity when only a subset of the population responds to treatments. These ranks may also be used in a stratified statistic for a block design [29, §2], and Renyi's result remains applicable, so transformed ranks may be used to construct evidence factors.

Despite this, many evidence factors cannot be derived from Renyi's result. Some evidence factors combine comparisons within pairs or blocks and comparisons between pairs or blocks; see Sect. 20.3 and [30]. Renyi's result does not apply in this situation. Is there another approach?

Renyi's result focuses attention on partial ranks, as in Table 20.2, but this focus greatly restricts the choice of test statistics. Perhaps we do not wish to use the partial ranks as the basis for the test. The stratified Wilcoxon statistic used to compare E2 in {E1,E2} in Table 20.1 uses partial ranks of 1 or 2 in each matched pair, so it is effectively the same as the sign test statistic for matched pairs. If the E2-minus-E1 pair differences in Table 20.1 were independent observations from the Normal distribution with expectation τ and variance 1, then in a randomized trial, Wilcoxon's signed rank test would have Pitman efficiency $3/\pi = 0.95$ relative to the t-test, whereas the sign test would have efficiency $2/\pi = 0.64$; so, by that standard, Wilcoxon's signed rank test would be a better choice of test than the sign test [8]. However, Wilcoxon's signed rank statistic is not a function of the partial ranks, so Renyi's result does not apply to it. In a sensitivity analysis with $\tau = 0.5$, we saw in Sect. 19.1 that Wilcoxon's signed rank statistic has design sensitivity $\widetilde{\Gamma} = 3.17$; however, in this same situation, the sign test has design sensitivity $\widetilde{\Gamma} = 2.24$; so, the sign test exaggerates the damage that an unmeasured bias of $\Gamma = 2.5$ can do [33, Table 2]. Moreover, other statistics have larger design sensitivity, $\widetilde{\Gamma}$, than does Wilcoxon's signed rank statistic; see Sect. 19.1 and [32]. We prefer to choose a test statistic based on its theoretical performance, rather than being confined to using partial ranks. Is it possible to have evidence factors without being confined to using partial ranks as the basis for a rank test?

Statistics that use partial ranks are a subset of the family of rank statistics. Many statistics, such as means, proportions, and M-statistics, do not use ranks of any kind, yet they can be the basis for randomization tests and sensitivity analyses; see, for instance, Sect. 2.8 and [28, 37]. Are there evidence factors for these statistics that do not involve ranks of any kind?

Most importantly, Renyi's result describes two factors that are each randomized. In Table 20.1, there is little reason to believe either factor is randomized. There is a sensitivity analysis for each factor taken separately, focusing on nonrandom

treatment assignment for that factor alone. When treatments are not randomly assigned, when sensitivity analyses are needed for each factor, ask: Does Table 20.1 yield two essentially independent sensitivity analyses? If so, can these two separate sensitivity analyses be combined as if they were from independent studies?

The questions raised in this section are answered in the slightly technical appendix Sect. 20.4 based on [31, 36], but first the answers are illustrated in Sect. 20.3.

20.2.6 Sensitivity Analyses for Evidence Factors

As a consequence of Proposition 20.2 in Sect. 20.4, we may conduct a sensitivity analysis for the two stratified rank sum tests applied to tannery workers example [44] in Table 20.1. As we saw, both factors, C to {E1, E2, C} and E2 to {E1, E2}, yield small P-values when testing the null hypothesis H_0 of no effect in a randomization test, $\Gamma = 1$. The first factor, C to {E1, E2, C}, becomes sensitive at about $\Gamma = 6$, with an upper bound on its one-sided P-value of 0.0505. The second factor, E2 to {E1, E2}, becomes sensitive at about $\Gamma = 1.7$ with an upper bound on its one-sided P-value of 0.0601. Fisher's method of combining two independent P-values uses, in effect, the sampling distribution of the product of two independent uniform random variables. Combining the two P-value bounds, 0.0505 and 0.0601, in this way gives a combined P-value of 0.0206, so bias of $\Gamma = 6$ in the first factor together with bias of $\Gamma = 1.7$ in the second factor could just barely lead to acceptance of H_0 based on either factor alone, but leads to rejection of H_0 at the 0.05 level using the two factors together.

Fisher's method tends to be conservative when applied to sensitivity analyses. Fisher's method assumes that P-values are uniformly distributed under the null hypothesis, but for large enough Γ an upper bound on a P-value is likely to be fixed close to 1. Jesse Hsu and colleagues [10] show that sensitivity analyses have higher power if P-values are combined using the truncated product method of Dmitri Zaykin, Lev Zhivotovsky, Peter Westfall, and Bruce Weir [43], in which the combination statistic is the product of those P-values that are less than or equal to some truncation point, `trunc`, perhaps 0.1 or 0.2. Fisher's method is the same as setting `trunc` to 1. Where Fisher's method combined P-values of 0.0505 and 0.0601 to 0.0206, truncation at 0.1 gives a combined P-value of 0.012.

As noted earlier, the stratified Wilcoxon test becomes the sign test in the comparison of E2 to {E1, E2}, and the sign test has poor Pitman efficiency in a randomization test, $\Gamma = 1$, and poor design sensitivity in a sensitivity analysis with $\Gamma > 1$. Suppose, instead, we use the randomization distribution of Huber's m-statistic [11, 21, 22, 28], with Huber's ψ-function as defined in Sect. 2.8; see also Sect. 19.2. [14]. In this case, strict independence of the two factors is lost, but Proposition 20.2 shows that the analysis above remains valid if m-statistics replace stratified Wilcoxon statistics.

Using the permutation distribution of the Huber–Maritz m-statistic to test H_0, comparing C to {E1, E2, C}, the upper bound on the one-sided P-value is 0.049 at $\Gamma = 11.7$, so the first factor is insensitive to much larger biases if the m-statistic is used in place of the stratified Wilcoxon statistic. In the second factor, comparing E2 to {E1, E2}, the upper bound on the one-sided P-value is 0.035 at $\Gamma = 2$, so again the m-statistic reports insensitivity to larger biases than does the stratified Wilcoxon statistic. Combining the two P-values using the truncated product, truncating at 0.1, yields a combined P-value of 0.0079.

For details of the computations discussed in this section, see Sect. 20.7 and [31, 34].

20.3 A Second Example: Smoking and Periodontal Disease

Recall Fig. 19.2 in the study of smoking as a possible cause of periodontal disease in Sect. 19.4. Panel (i) of Fig. 19.2 depicts smoker-minus-control pair differences, Y_i, in a measure of periodontal disease, ignoring the amount smoked, d_i, by the smoker. Panel (ii) of Fig. 19.2 asked whether larger Y_i tend to occur when the smoker smoked more cigarettes each day, d_i. Intuitively, these seem to be two different questions: whether Panel (i) is tilted toward positive Y_i seems quite separate from whether larger Y_i occur with larger d_i in Panel (ii). Indeed, this intuition is correct, and Panels (i) and (ii) provide two evidence factors.

As discussed in Sect. 20.4 and [30, 36], two evidence factors are formed by applying a signed rank statistic to Panel (i) of Fig. 19.2 and the crosscut statistic to Panel (ii). Table 20.4 applies the U-statistic (8,6,8) from Sect. 19.2 to Panel (i) of Fig. 19.2 and the crosscut statistic from Sect. 19.4 to Panel (ii), setting $\eta = 0.2$.

Table 20.4 Evidence factor analysis for smoking as a cause of periodontal disease. The columns refer to the U-statistic (8, 6, 8) applied to Panel (i) of Fig. 19.2 with sensitivity parameter Γ_1. The rows refer to the crosscut statistic with $\eta = 0.2$ applied to Panel (ii) of Fig. 19.2 and sensitivity parameter Γ_2. The first row of P-value bounds is for the U-statistic used alone, and the first column of P-value bounds is for the crosscut statistic used alone. The interior of the table combines the U-statistic bound and the crosscut statistic bound using the truncated product of P-values, truncating at 0.1

		Γ_1 for U-statistic					
Γ_2 for		1.0	3.0	3.3	3.5	3.7	∞
Crosscut	Crosscut ↓; U →	0.0000	0.0030	0.0136	0.0304	0.0591	1.0000
1	0.0022	0.0000	0.0001	0.0003	0.0005	0.0009	0.0095
1.4	0.0214	0.0000	0.0005	0.0018	0.0036	0.0061	0.0484
1.6	0.0443	0.0000	0.0009	0.0034	0.0065	0.0108	0.0898
1.8	0.0781	0.0000	0.0015	0.0054	0.0101	0.0165	0.1507
∞	1.0000	0.0000	0.0119	0.0345	0.0647	0.1163	1.0000

The first row of bounds on P-values refer to the U-statistic used alone. The first column of bounds on P-values refers to the crosscut statistic used alone. The interior of the table combines these two P-values using the truncated product method, with truncation at 0.1. The sensitivity parameter Γ_1, for the U-statistic, describes potential bias in who smokes in Panel (i) of Fig. 19.2, while Γ_2, for the crosscut statistic, describes potential bias in how many cigarettes smokers smoke in Panel (ii).

Bikram Karmakar and colleagues [13] suggest that closed testing [19] should be used to interpret two P-values from two evidence factors. Specifically, at (Γ_1, Γ_2), they do not reject the hypothesis H_0 of no effect at level α, conventionally $\alpha = 0.05$, if the P-value from the truncated product is greater than α; otherwise, they reject H_0, saying that rejection is corroborated by each factor whose marginal P-value bound is at most α. For instance, in a 0.05 level test at $(\Gamma_1, \Gamma_2) = (3.5, 1.6)$, H_0 is rejected with combined P-value bound of 0.0065, and each factor corroborates this rejection. In contrast, at $(\Gamma_1, \Gamma_2) = (3.7, 1.8)$, H_0 is rejected with combined P-value bound of 0.0165, but neither factor leads to rejection on its own. Karmakar and colleagues [13] discuss the Bahadur efficiency of this method. Their method becomes more complex with three evidence factors.

Notably in Table 20.4, even $\Gamma_1 = \infty$ leads to rejection of H_0 providing $\Gamma_2 \leq 1.4$ as the combined P-value is at most 0.0484 in the final column of Table 20.4. Also, even $\Gamma_2 = \infty$ leads to rejection of H_0 providing $\Gamma_1 \leq 3.3$ as the combined P-value is at most 0.0345 in the final row of Table 20.4. Complete invalidation of either factor does not lead to acceptance of H_0 providing the bias affecting the other factor is not too large.

20.4 Appendix: Some Theory

20.4.1 A Tiny Example

This appendix provides some general theory for two evidence factors, simplifying certain arguments from [36]. The goal is Proposition 20.2, which provides simple, sufficient conditions for two sensitivity analysis to constitute two essentially independent evidence factors.

Consider an experiment involving N individuals, $1, \ldots, N$, whose N labels are placed in a column vector $\mathbf{n} = (1, 2, \ldots, N)^T$. The experiment also has N treatment positions. A treatment assignment simply means assigning each individual to a treatment position. For instance, if $N = 4$, then $(3, 4, 2, 1)^T$ assigns individual 3 to the first treatment position, individual 4 to the second treatment position, and so on.

The example discussed throughout this section consists of two pairs from the observational study of smoking and periodontal disease discussed in Sect. 19.4 and depicted in Fig. 19.2. With two pairs, we can inspect the relevant 4×4 permutation matrices, but a parallel discussion applies to Fig. 19.2 as a whole,

where the permutation matrices would be $2I \times 2I = 882 \times 882$ matrices. The discussion that follows permutes two doses among the two pairs, and makes one smoker-control treatment assignment in each of two pairs, so there are $2 \times 2^2 = 8$ permutation matrices to be considered. In Fig. 19.2, there are $I! \times 2^I = 441! \times 2^{441}$ permutation matrices to consider; however, no new issues arise as I moves from $I = 2$ to $I = 441$.

Table 20.5 describes a paired experiment with $N = 4$ treatment positions. Individuals were paired before treatment based on covariates, and we consider only treatment assignments that retain this pairing, so we do not consider all $N! = 4! = 4 \times 3 \times 2 \times 1 = 24$ treatment assignments, but rather a subset of them. Each pair in Table 20.5 will contain one smoker and one nonsmoking control. In pair 1, the smoker will smoke 40 cigarettes per day. In pair 2, the smoker will smoke 8 cigarettes per day.

Table 20.6 describes $N = 4$ individuals in two pairs of two individuals. They are four individuals from the nonrandomized observational study described in Sect. 19.4 and depicted in Fig. 19.2. The pairs were matched for gender, age, black race, education, and income. The first pair contains two men in their 50s with some college and family incomes at least equal to five times the poverty level. The second pair contains two women in their 60s with a high school or equivalent degree and family incomes below the poverty level.

Again, we will assign the four individuals in Table 20.6 to the four positions in Table 20.5 without breaking up the pairs. There are $8 = 2 \times 2^2$ ways to do this: assign either pair 1 or pair 2 in Table 20.6 to pair 1 in Table 20.5; then, assign one individual to smoking or control within each pair in Table 20.5. The analogous situation in Fig. 19.2 would entail 441! permutations of doses, d_i, among the 441

Table 20.5 Four treatment positions in two matched pairs

Structure		Treatment	
Position	Pair	Smoker	Cigarettes
1	1	1	40
2	1	0	0
3	2	1	8
4	2	0	0

Table 20.6 Four individuals in two matched pairs, matched for gender (1 = female), age in years, black or other race (1 = black), education level, and income recorded as the ratio of family income to the poverty level, capped at 5 × poverty

Individual		Covariates				
ID		Matched covariates				
i	Pair	Female	Age	Black	Education	Income
1	1	0	54	0	Some college	5.00
2	1	0	53	0	Some college	5.00
3	2	1	61	1	HS/GED	0.67
4	2	1	64	1	HS/GED	0.72

Table 20.7 Outcomes for four individuals in two matched pairs. The outcome is the percent of dental measurements indicative of periodontal disease

Individual		Outcome
i	Pair	Percent diseased
1	1	63.04
2	1	00.00
3	2	21.50
4	2	10.42

pairs, and 2^{441} treatment assignments within pairs, making $441! \times 2^{441}$ treatment assignments in total.

As it turns out, in the actual study, the four individuals in Table 20.6 were assigned in their current order, $(1, 2, 3, 4)$, to the four treatment positions in Table 20.5. As it turns out, in Table 20.7, individual 1 had periodontal disease in 63.04% of his measures of tooth and gum separation, individual 2 had 0%, individual 3 had 21.5%, and individual 4 had 10.42%. This result provides two associations consistent with smoking causing an increase in periodontal disease. First, in both pairs, the smoker had greater periodontal disease than the matched control, $63.04 - 0.00 = 63.04 > 0$ and $21.50 - 10.42 = 11.08 > 0$. Second, the treated-minus-control difference was larger in the pair in which 40 cigarettes were smoked by the smoker than in the pair where 8 cigarettes were smoked by the smoker, $63.04 > 11.08$. We might have seen either association without the other, so their joint occurrence is news. Had we applied Wilcoxon's signed rank test to the two matched pair differences, it would have ignored the amount smoked by the two smokers, so it is ignoring the information provided by the second association. In a different sense, had we applied Kendall's correlation to relate the amounts smoked, $(40, 8)$, to the differences in periodontal disease, $(63.04, 11.08)$, it would have ignored the information important to Wilcoxon's test. Indeed, if we had randomly picked one of the $8 = 2 \times 2^2$ possible treatment assignments, and if Fisher's hypothesis of no treatment effect were true, then Wilcoxon's signed rank statistic would be independent of Kendall's correlation; see [30, §5.1]. Despite looking twice at the same data about the same people, it is as if we had done two independent studies on unrelated people. When there are evidence factors, a single study can provide an independent replication of itself. The goal in this appendix is to understand issues of this sort in greater generality.

An $N \times N$ permutation matrix, $\mathbf{g}$, is a matrix of 0's and 1's such that each row contains a single 1 and each column contains a single 1. If $\mathbf{n} = (1, 2, \ldots, N)^T$, then $\mathbf{gn}$ is a permutation of the numbers $1, \ldots, N$. For instance, the permutation matrix $\mathbf{g}$ in (20.1),

$$\mathbf{g} = \begin{bmatrix} 0\,0\,1\,0 \\ 0\,0\,0\,1 \\ 0\,1\,0\,0 \\ 1\,0\,0\,0 \end{bmatrix}, \tag{20.1}$$

permutes $\mathbf{n} = (1, 2, 3, 4)^T$ to yield the treatment assignment $\mathbf{gn} = (3, 4, 2, 1)^T$. This treatment assignment preserves the pairing in Table 20.6, assigning individual 3 to smoking 40 cigarettes per day, individual 4 as control in the same pair, individual 2 to smoking 8 cigarettes per day, individual 1 as control in the same pair.

In (20.2), $\mathfrak{H}$ contains $2 = |\mathfrak{H}|$ permutation matrices, namely $\mathbf{h}_1$ and $\mathbf{h}_2$.

$$\mathfrak{H} = \left\{ \mathbf{h}_1 = \begin{bmatrix} 1\,0\,0\,0 \\ 0\,1\,0\,0 \\ 0\,0\,1\,0 \\ 0\,0\,0\,1 \end{bmatrix}, \mathbf{h}_2 = \begin{bmatrix} 0\,0\,1\,0 \\ 0\,0\,0\,1 \\ 1\,0\,0\,0 \\ 0\,1\,0\,0 \end{bmatrix} \right\} \tag{20.2}$$

Both $\mathbf{h}_1$ and $\mathbf{h}_2$ permute $\mathbf{n} = (1, 2, 3, 4)^T$ to yield treatment assignments that preserve the pairing in Table 20.6, namely $\mathbf{h}_1\mathbf{n} = (1, 2, 3, 4)^T$ and $\mathbf{h}_2\mathbf{n} = (3, 4, 1, 2)^T$.

20.4.2 Elementary Theory of Groups of Permutation Matrices

A set $\mathfrak{G}$ of $N \times N$ permutation matrices is a group if: (i) it contains the identity matrix, (ii) the product of any two matrices in $\mathfrak{G}$ is also in $\mathfrak{G}$, and (iii) the inverse of any matrix in $\mathfrak{G}$ is also in $\mathfrak{G}$. Is $\mathfrak{H}$ in (20.2) a group? First, $\mathbf{h}_1 \in \mathfrak{H}$ is the identity matrix, $\mathbf{I}$. Next, check the products of two matrices to see if they are in $\mathfrak{H}$: $\mathbf{h}_1\mathbf{h}_1 = \mathbf{h}_1 \in \mathfrak{H}$, $\mathbf{h}_2\mathbf{h}_2 = \mathbf{h}_1 \in \mathfrak{H}$, $\mathbf{h}_1\mathbf{h}_2 = \mathbf{h}_2\mathbf{h}_1 = \mathbf{h}_2 \in \mathfrak{H}$. Finally, check to see if the inverses of matrices in $\mathfrak{H}$ are also in $\mathfrak{H}$: because $\mathbf{h}_2\mathbf{h}_2 = \mathbf{I}$, we have $\mathbf{h}_2^{-1} = \mathbf{h}_2 \in \mathfrak{H}$, and of course $\mathbf{h}_1^{-1} = \mathbf{h}_1 \in \mathfrak{H}$. So $\mathfrak{H}$ is a group of permutation matrices.

In the same way, it is easy to check that $\mathfrak{K}$ in (20.3) is a group of permutation matrices,

$$\mathfrak{K} = \{\mathbf{k}_1, \mathbf{k}_2, \mathbf{k}_3, \mathbf{k}_4\}, \tag{20.3}$$

where

$$\mathbf{k}_1 = \begin{bmatrix} 1\,0\,0\,0 \\ 0\,1\,0\,0 \\ 0\,0\,1\,0 \\ 0\,0\,0\,1 \end{bmatrix}, \quad \mathbf{k}_2 = \begin{bmatrix} 0\,1\,0\,0 \\ 1\,0\,0\,0 \\ 0\,0\,1\,0 \\ 0\,0\,0\,1 \end{bmatrix},$$

$$\mathbf{k}_3 = \begin{bmatrix} 1\,0\,0\,0 \\ 0\,1\,0\,0 \\ 0\,0\,0\,1 \\ 0\,0\,1\,0 \end{bmatrix}, \quad \mathbf{k}_4 = \begin{bmatrix} 0\,1\,0\,0 \\ 1\,0\,0\,0 \\ 0\,0\,0\,1 \\ 0\,0\,1\,0 \end{bmatrix}.$$

Moreover, $\mathbf{k}_1\mathbf{n} = (1, 2, 3, 4)^T$, $\mathbf{k}_2\mathbf{n} = (2, 1, 3, 4)^T$, $\mathbf{k}_3\mathbf{n} = (1, 2, 4, 3)^T$, and $\mathbf{k}_4\mathbf{n} = (2, 1, 4, 3)^T$ are each treatment assignments that preserve the pairing in Table 20.6. Moreover, if we multiply two permutation matrices that preserve the pairing in Table 20.6, then their product is a permutation matrix that preserves the pairing in Table 20.6. For instance, $\mathbf{k}_2\mathbf{k}_4$ preserves the pairing in Table 20.6.

For two sets, $\mathfrak{H}$ and $\mathfrak{K}$, of $N \times N$ permutation matrices, write $\mathfrak{H}\mathfrak{K}$ for the set of products of the form $\{\mathbf{hk} : \mathbf{h} \in \mathfrak{H},\ \mathbf{k} \in \mathfrak{K}\}$. In (20.2) and (20.3), the eight products, $\mathbf{hk}$ with $\mathbf{h} \in \mathfrak{H}$ and $\mathbf{k} \in \mathfrak{K}$ are all distinct, and they form the $8 = 2 \times 2^2$ ways to assign the individuals in Table 20.6 to the treatment positions in Table 20.5.

Notice that $\mathbf{h}_2\mathbf{k}_2 \neq \mathbf{k}_2\mathbf{h}_2$, so the matrices in $\mathfrak{H}\mathfrak{K}$ do not generally commute. In fact, $\mathbf{h}_2\mathbf{k}_2 = \mathbf{k}_3\mathbf{h}_2$. Also, $\mathfrak{H}\mathfrak{K}$ contains only 8 permutation matrices, not the full set of $4! = 24$ permutation matrices of size 4×4. Nonetheless, perhaps surprisingly, $\mathfrak{H}\mathfrak{K} = \mathfrak{K}\mathfrak{H}$ or equivalently $\{\mathbf{hk} : \mathbf{h} \in \mathfrak{H},\ \mathbf{k} \in \mathfrak{K}\} = \{\mathbf{kh} : \mathbf{h} \in \mathfrak{H},\ \mathbf{k} \in \mathfrak{K}\}$, so both $\mathfrak{H}\mathfrak{K}$ and $\mathfrak{K}\mathfrak{H}$ contain the 8 permutation matrices that preserve the pairs in Table 20.6.

With a bit of patience, in (20.2) and (20.3) we may check that $\mathfrak{H}\mathfrak{K}$ is itself a group of permutation matrices. However, there is a less tedious way to check that $\mathfrak{H}\mathfrak{K}$ is a group. Indeed, Isaacs [12, Lemma 2.18, p. 22] demonstrates:

Proposition 20.1 *Let $\mathfrak{H}$ and $\mathfrak{K}$ be two groups of $N \times N$ permutation matrices. Then $\mathfrak{H}\mathfrak{K}$ is a group of $N \times N$ permutation matrices if and only if $\mathfrak{H}\mathfrak{K} = \mathfrak{K}\mathfrak{H}$.*

Let $\mathfrak{G}$ be a group of $N \times N$ permutation matrices. A subset $\mathfrak{H} \subseteq \mathfrak{G}$ is a subgroup if $\mathfrak{H}$ is also a group. For instance, $\mathfrak{H}$ in (20.2) is a subgroup of $\mathfrak{G} = \mathfrak{H}\mathfrak{K}$, and $\mathfrak{K}$ in (20.3) is a subgroup of $\mathfrak{G} = \mathfrak{H}\mathfrak{K}$; moreover, $\{\mathbf{k}_1\} = \{\mathbf{I}\} \subseteq \mathfrak{G}$ is a subgroup of $\mathfrak{G}$, $\{\mathbf{k}_1, \mathbf{k}_2\} \subseteq \mathfrak{K} \subseteq \mathfrak{G}$ is a subgroup of both $\mathfrak{K}$ and $\mathfrak{G}$, and trivially $\mathfrak{G}$ is a subgroup of $\mathfrak{G}$. However, $\{\mathbf{k}_2\} \subseteq \mathfrak{K} \subseteq \mathfrak{G}$ is not a subgroup because $\mathbf{k}_2\mathbf{k}_2 = \mathbf{I} \notin \{\mathbf{k}_2\}$.

Let $\mathfrak{G}$ be a group of $N \times N$ permutation matrices, and let $\mathfrak{H} \subseteq \mathfrak{G}$ be a subgroup. For any permutation matrix $\mathbf{g} \in \mathfrak{G}$, abbreviate $\mathfrak{H}\{\mathbf{g}\} = \{\mathbf{hg} : \mathbf{h} \in \mathfrak{H},\ \mathbf{g}\}$ by $\mathfrak{H}\mathbf{g}$. The set $\mathfrak{H}\mathbf{g} \subseteq \mathfrak{G}$ is called a coset of $\mathfrak{H}$. For example, $\mathfrak{H}$ in (20.2) is the coset $\mathfrak{H}\mathbf{I} \subseteq \mathfrak{G} = \mathfrak{H}\mathfrak{K}$, and $\mathfrak{H}\mathbf{k}_2$ is the coset

$$\mathfrak{H}\mathbf{k}_2 = \{\mathbf{k}_2,\ \mathbf{h}_2\mathbf{k}_2\} = \left\{ \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} \right\}.$$

For a fixed $N \times N$ permutation matrix $\mathbf{g} \in \mathfrak{G}$, the elements $\mathbf{hg}$, $\mathbf{h} \in \mathfrak{H}$, are all different (because $\mathbf{g}$ has an inverse), so every coset contains the same number of $N \times N$ permutation matrices, $|\mathfrak{H}\mathbf{g}| = |\mathfrak{H}|$. Each of the cosets of $\mathfrak{H} \subset \mathfrak{G} = \mathfrak{H}\mathfrak{K}$ in (20.2) and (20.3) contains two 4×4 permutation matrices. If $\mathbf{g}_1, \mathbf{g}_2 \in \mathfrak{G}$, then $\mathfrak{H}\mathbf{g}_1$ and $\mathfrak{H}\mathbf{g}_2$ are the same coset, $\mathfrak{H}\mathbf{g}_1 = \mathfrak{H}\mathbf{g}_2$, if and only if $\mathfrak{H}\mathbf{g}_1\mathbf{g}_2^{-1} = \mathfrak{H}$, and this occurs if and only if $\mathbf{g}_1\mathbf{g}_2^{-1} \in \mathfrak{H}$. Repeatedly using the fact that $\mathfrak{H}$ is a group, the condition $\mathbf{g}_1\mathbf{g}_2^{-1} \in \mathfrak{H}$ is seen, as follows, to define an equivalence relation $\simeq$ on $\mathfrak{G}$: (i) $\mathbf{gg}^{-1} = \mathbf{I} \in \mathfrak{H}$, so $\mathbf{g} \simeq \mathbf{g}$ for each $\mathbf{g} \in \mathfrak{G}$, (ii) $\mathbf{g}_1\mathbf{g}_2^{-1} \in \mathfrak{H}$ implies $\left(\mathbf{g}_1\mathbf{g}_2^{-1}\right)^{-1} =$

$\mathbf{g}_2\mathbf{g}_1^{-1} \in \mathfrak{H}$, so $\mathbf{g}_1 \simeq \mathbf{g}_2$ implies $\mathbf{g}_2 \simeq \mathbf{g}_1$, (iii) $\mathbf{g}_1\mathbf{g}_2^{-1} \in \mathfrak{H}$ and $\mathbf{g}_2\mathbf{g}_3^{-1} \in \mathfrak{H}$ jointly imply $\mathbf{g}_1\mathbf{g}_3^{-1} = \mathbf{g}_1\mathbf{g}_2^{-1}\mathbf{g}_2\mathbf{g}_3^{-1} \in \mathfrak{H}$, so $\mathbf{g}_1 \simeq \mathbf{g}_2$ and $\mathbf{g}_2 \simeq \mathbf{g}_3$ jointly imply $\mathbf{g}_1 \simeq \mathbf{g}_3$. So the cosets of $\mathfrak{H} \subseteq \mathfrak{G}$ partition $\mathfrak{G}$ into mutually exclusive and exhaustive equivalence classes, each class or coset containing $|\mathfrak{H}|$ permutation matrices, making $|\mathfrak{G}| / |\mathfrak{H}|$ distinct cosets. For $\mathfrak{H} \subset \mathfrak{G} = \mathfrak{H}\mathfrak{K}$ in (20.2) and (20.3), the $|\mathfrak{G}| = 8$ permutation matrices partition into $|\mathfrak{G}| / |\mathfrak{H}| = 8/2 = 4$ cosets $|\mathfrak{H}\mathbf{g}|$, each containing $|\mathfrak{H}| = 2$ permutation matrices.

As just noted, for a subgroup $\mathfrak{H}$ of a group $\mathfrak{G}$ of $N \times N$ permutation matrices, we have $L = |\mathfrak{G}| / |\mathfrak{H}|$ distinct cosets, but alas they have a total of $|\mathfrak{G}|$ different names, namely $\mathfrak{H}\mathbf{g}$ for $\mathbf{g} \in \mathfrak{G}$. It is convenient to have one name for each coset, not many different names for each coset. A transversal $\mathfrak{T} = \{\mathbf{k}_1, \ldots, \mathbf{k}_L\} \subseteq \mathfrak{G}$ is a set of $L = |\mathfrak{G}| / |\mathfrak{H}|$ permutation matrices such that each $\mathfrak{H}\mathbf{k}_\ell$ is a different coset, with the consequence that each coset has a different name, namely $\mathfrak{H}\mathbf{k}_\ell$; then, $\mathfrak{G}$ is the disjoint union of the $\mathfrak{H}\mathbf{k}_\ell$, $\mathfrak{G} = \mathfrak{H}\mathbf{k}_1 \cup \cdots \cup \mathfrak{H}\mathbf{k}_L$ with $\emptyset = \mathfrak{H}\mathbf{k}_\ell \cap \mathfrak{H}\mathbf{k}_{\ell'}$ if $\ell \neq \ell'$ and $|\mathfrak{H}\mathbf{k}_\ell| = |\mathfrak{H}|$ for each ℓ. Thus, each $\mathbf{g} \in \mathfrak{G}$ can be represented in exactly one way as the product of an element of $\mathfrak{H}$ and an element of $\mathfrak{T}$, that is, as $\mathbf{g} = \mathbf{hk}$ with $\mathbf{h} \in \mathfrak{H}$ and $\mathbf{k} \in \mathfrak{T}$. It is easy to construct a transversal: list the distinct cosets and pick an arbitrary element $\mathbf{k}_\ell$ from the ℓth coset. For review of groups, subgroups, cosets, and transversals, see Kurzweil and Stellmacher [17, §1.1].

In general, a transversal need not be a subgroup. However, $\mathfrak{K}$ in (20.3) happens to be both a subgroup of $\mathfrak{G} = \mathfrak{H}\mathfrak{K}$ and also a transversal of the $4 = |\mathfrak{K}| = |\mathfrak{G}| / |\mathfrak{H}| = 8/2$ cosets of $\mathfrak{H}$ in (20.2).

20.4.3 Probability Distributions on a Group of Permutation Matrices

A probability distribution on a group $\mathfrak{G} = \left\{\mathbf{g}_1, \ldots, \mathbf{g}_{|\mathfrak{G}|}\right\}$ of $N \times N$ permutation matrices is simply a vector $\mathbf{p} = \left(p_{\mathbf{g}_1}, \ldots, p_{\mathbf{g}_{|\mathfrak{G}|}}\right)^T$ of dimension $|\mathfrak{G}|$ with $p_\mathbf{g} \geq 0$ for each $\mathbf{g} \in \mathfrak{G}$ and $1 = \sum_{\mathbf{g}\in\mathfrak{G}} p_\mathbf{g}$. Here, $p_\mathbf{g}$ is the probability that we select $\mathbf{g} \in \mathfrak{G}$.

Consider a subgroup $\mathfrak{H}$ of $\mathfrak{G}$ and a transversal $\mathfrak{T} = \{\mathbf{k}_1, \ldots, \mathbf{k}_L\} \subseteq \mathfrak{G}$. Recall that each $\mathbf{g} \in \mathfrak{G}$ can be written in precisely one way as $\mathbf{g} = \mathbf{hk}$ with $\mathbf{h} \in \mathfrak{H}$ and $\mathbf{k} \in \mathfrak{T}$, so that sampling a $\mathbf{g} \in \mathfrak{G}$ entails sampling a unique $\mathbf{h} \in \mathfrak{H}$ and a unique $\mathbf{k} \in \mathfrak{T}$ such that $\mathbf{g} = \mathbf{hk}$. As a consequence, a distribution $\mathbf{p}$ on $\mathfrak{G}$ yields a marginal distribution on $\mathfrak{T}$ and a conditional distribution on $\mathfrak{H}$ given the value on $\mathfrak{T}$. The marginal distribution on $\mathfrak{T}$ attaches to $\mathbf{k} \in \mathfrak{T}$ the total probability of all $\mathbf{g} = \mathbf{hk}$ with $\mathbf{k} \in \mathfrak{T}$ fixed as $\mathbf{h}$ varies over $\mathfrak{H}$, so the marginal probability of $\mathbf{k} \in \mathfrak{T}$ is the total probability of all $\mathbf{g} = \mathbf{hk}$ in the coset $\mathfrak{H}\mathbf{k}$. If the $|\mathfrak{G}|$-dimensional distribution on $\mathfrak{G}$ is $\mathbf{p} = \left(p_1, \ldots, p_{|\mathfrak{G}|}\right)^T$, then the L-dimensional vector $\overline{\mathbf{p}} = \left(\overline{p}_{\mathbf{k}_1}, \ldots, \overline{p}_{\mathbf{k}_L}\right)^T$ that defines the marginal distribution on $\mathfrak{T}$ attaches probability $\overline{p}_\mathbf{k}$ to $\mathbf{k}$, where $\overline{p}_\mathbf{k} = \sum_{\mathbf{g}\in\mathfrak{H}\mathbf{k}} p_\mathbf{g}$. Also, the conditional distribution $\widetilde{\mathbf{p}}_\mathbf{k} = \left(\widetilde{p}_{\mathbf{h}_1}, \ldots, \widetilde{p}_{\mathbf{h}_{|\mathfrak{H}|}}\right)^T$ of $\mathbf{h} \in \mathfrak{H}$ given the value of $\mathbf{k} \in \mathfrak{T}$ is given by $p_\mathbf{g}/\overline{p}_\mathbf{k} = p_\mathbf{hk}/\overline{p}_\mathbf{k}$, where $\mathbf{g} = \mathbf{hk}$, providing $\overline{p}_\mathbf{k} > 0$.

Suppose that we assign treatments by picking $\mathbf{g} \in \mathfrak{G}$ with probability $p_{\mathbf{g}}$, and placing individuals $\mathbf{n} = (1, 2, \ldots, N)^T$ in treatment positions $\mathbf{gn}$. Then this is the same as picking $\mathbf{k} \in \mathfrak{T}$ with probability $\overline{p}_{\mathbf{k}} = \sum_{\mathbf{g} \in \mathfrak{H}\mathbf{k}} p_{\mathbf{g}}$, picking $\mathbf{h} \in \mathfrak{H}$ with conditional probability $p_{\mathbf{g}}/\overline{p}_{\mathbf{k}} = p_{\mathbf{hk}}/\overline{p}_{\mathbf{k}}$, and placing individuals $\mathbf{n} = (1, 2, \ldots, N)^T$ in treatment position $\mathbf{gn} = \mathbf{hkn}$. In a completely randomized experiment in Tables 20.5 and 20.6, we might pick $\mathbf{k}$ at random in (20.3), so $\overline{p}_{\mathbf{k}} = 1/4$ for each $\mathbf{k}$ or $\overline{\mathbf{p}} = (1/4,\ 1/4,\ 1/4,\ 1/4)^T$ and $\mathbf{h} \in \mathfrak{H}$ at random in (20.2), so $\widetilde{\mathbf{p}}_{\mathbf{k}} = (1/2,\ 1/2)^T$ for each $\mathbf{k}$; then $p_{\mathbf{g}} = 1/8$ and $\mathbf{p} = (1/8, \ldots, 1/8)^T$. It will be important later that $\overline{\mathbf{p}}$ could be completely randomized while $\widetilde{\mathbf{p}}_{\mathbf{k}}$ departs from complete randomization, or $\widetilde{\mathbf{p}}_{\mathbf{k}}$ could be completely randomized given $\mathbf{k}$ while $\overline{p}_{\mathbf{k}}$ departs from complete randomization, and in either case $\mathbf{p}$ departs from complete randomization. Also, the two components, $\overline{\mathbf{p}}$ and $\widetilde{\mathbf{p}}_{\mathbf{k}}$, might depart from completely random assignment to different quantitative degrees.

20.4.4 Invariant Test Statistics

Table 20.7 shows the periodontal outcomes for the four individuals in Table 20.6. If smoking had no effect on the teeth of these individuals, then by the definition of no effect, swapping their treatment positions in Table 20.5 would not alter their outcomes. So, under Fisher's null hypothesis, H_0, of no treatment effect, permuting the four individuals in Table 20.7 to positions in Table 20.5 permutes the outcomes among treatment positions but does not change the outcomes themselves. In contrast, if smoking did have an effect, perhaps whoever was placed in position 1 in Table 20.5, smoking 40 cigarettes every day, would have the most extensive periodontal disease.

This appendix discusses testing Fisher's hypothesis, H_0, and the associated null distributions of test statistics, so Fisher's H_0 is assumed to be true. When H_0 is true, a test statistic T is a function of the treatment assignment $\mathbf{g} \in \mathfrak{G}$ and the unchanging outcomes in Table 20.7, so we may write $T = t(\mathbf{g})$.

For a subgroup $\mathfrak{H}$ of $\mathfrak{G}$ and a transversal $\mathfrak{T} = \{\mathbf{k}_1, \ldots, \mathbf{k}_L\} \subseteq \mathfrak{G}$ of its cosets, a statistic $T = t(\mathbf{g})$ is invariant with respect to $\mathfrak{H}$ if $t(\mathbf{hk}) = t(\mathbf{k})$ for all $\mathbf{h} \in \mathfrak{H}$ and $\mathbf{k} \in \mathfrak{T}$. For instance, Wilcoxon's signed rank statistic comparing smokers and matched controls in Tables 20.5 and 20.7 takes no account of the number of cigarettes smoked, so swapping the identities of the two pairs using $\mathbf{h} \in \mathfrak{H}$ does not change its value: it is invariant.

20.4.5 Ignoring One Evidence Factor

Typically, in observational studies, we do know how treatments were assigned. A sensitivity analysis considers a set of possible distributions of treatment assignments and asks whether there is at least one distribution in the set that might reasonably

have produced the observed association between treatment and outcome in the absence of a treatment effect. There is a subgroup $\mathfrak{H}$ of a group $\mathfrak{G}$ of treatment assignments, and a transversal $\mathfrak{T} = \{\mathbf{k}_1, \ldots, \mathbf{k}_L\} \subseteq \mathfrak{G}$ of the cosets of $\mathfrak{H}$.

Let $\mathscr{P}$ be a set of possible distributions $\mathbf{p} = \left(p_{\mathbf{g}_1}, \ldots, p_{\mathbf{g}_{|\mathfrak{G}|}}\right)^T$ of treatment assignments $\mathbf{g} \in \mathfrak{G}$. So $\mathscr{P}$ is a set whose elements are $|\mathfrak{G}|$-dimensional vectors, $\mathbf{p}$, each of which is a probability distribution, $p_{\mathbf{g}} \geq 0$ for each $\mathbf{g} \in \mathfrak{G}$, $1 = \sum_{\mathbf{g}\in\mathfrak{G}} p_{\mathbf{g}}$. We additionally require that the set $\mathscr{P}$ be closed; therefore, being closed and bounded, $\mathscr{P}$ is compact. Typically, we consider several such sets $\mathscr{P}$ that vary with the values of sensitivity parameters, but to keep the notation simple, let us consider just one set $\mathscr{P}$ corresponding with one specific value for sensitivity parameters. For each $\mathbf{p} \in \mathscr{P}$, there is a corresponding marginal distribution $\overline{p}_{\mathbf{k}} = \sum_{\mathbf{g}\in\mathfrak{H}\mathbf{k}} p_{\mathbf{g}}$ on $\mathfrak{T}$. Let $\mathscr{P}_{\mathfrak{T}}$ be the set of $|\mathfrak{T}|$-dimensional marginal distributions $\overline{\mathbf{p}}$ on $\mathfrak{T}$ such each $\overline{\mathbf{p}} \in \mathscr{P}_{\mathfrak{T}}$ is the marginal distribution corresponding with at least one $\mathbf{p} \in \mathscr{P}$. To emphasize, $\overline{\mathbf{p}} \in \mathscr{P}_{\mathfrak{T}}$ if and only if $\overline{\mathbf{p}}$ is the marginal distribution for at least one $\mathbf{p} \in \mathscr{P}$. Write $\chi(E) = 1$ if event E occurs and $\chi(E) = 0$ otherwise.

The upper bound on the tail probability that a statistic $t(\mathbf{g})$ is at least a for at least one $\mathbf{p} \in \mathscr{P}$ is

$$\max_{\mathbf{p}\in\mathscr{P}} \sum_{\mathbf{g}\in\mathfrak{G}} \chi\{t(\mathbf{g}) \geq a\}\ p_{\mathbf{g}} = \max_{\mathbf{p}\in\mathscr{P}} \sum_{\mathbf{h}\in\mathfrak{H}} \sum_{\mathbf{k}\in\mathfrak{T}} \chi\{t(\mathbf{hk}) \geq a\}\ p_{\mathbf{hk}}, \tag{20.4}$$

because each $\mathbf{g} \in \mathfrak{G}$ has a unique representation as $\mathbf{g} = \mathbf{hk}$ with $\mathbf{h} \in \mathfrak{H}$ and $\mathbf{k} \in \mathfrak{T}$. A minor technical point deserves brief comment. Expression (20.4) refers to a maximum, rather than a supremum: Is this justified? Viewed as a function of $\mathbf{p} \in \mathscr{P}$, the function being optimized in (20.4) is the sum of some of the coordinates of $\mathbf{p}$, so (20.4) is the optimum of a continuous function of $\mathbf{p}$ on a compact set $\mathscr{P}$; hence, the optimum is achieved in $\mathscr{P}$, so the use of maximum is justified.

Of course, the joint distribution $\mathbf{p} \in \mathscr{P}$ in (20.4) is not the usual basis for a sensitivity analysis that only involves $\mathbf{k}$, such as the use of Wilcoxon's signed rank statistic to compare daily smokers and never-smokers, ignoring the amount smoked. If we were ignoring the amount smoked, we would likely use the marginal distribution of $\mathbf{k}$. If indeed we simply ignored $\mathbf{h} \in \mathfrak{H}$, viewing it as fixed, performing the sensitivity analysis using $\mathbf{k} \in \mathfrak{T}$ with its corresponding marginal distributions $\overline{\mathbf{p}} \in \mathscr{P}_{\mathfrak{T}}$, the upper bound on the tail probability would be

$$\max_{\overline{\mathbf{p}}\in\mathscr{P}_{\mathfrak{T}}} \sum_{\mathbf{k}\in\mathfrak{T}} \chi\{t(\mathbf{k}) \geq a\}\ \overline{p}_{\mathbf{k}}. \tag{20.5}$$

Lemma 20.1 gives a natural condition on the test statistic, $t(\mathbf{g})$, such that (20.4) and (20.5) are equal.

Lemma 20.1 *If $t(\mathbf{g})$ is invariant with respect to $\mathfrak{H}$, then (20.4) and (20.5) are equal.*

Proof Recall that $\overline{\mathbf{p}} \in \mathscr{P}_{\mathfrak{T}}$ if and only if there exists a $\mathbf{p} \in \mathscr{P}$ such that $\overline{\mathbf{p}}$ is the marginal distribution of $\mathbf{k}$ for the joint distribution $\mathbf{p}$. Because $t(\cdot)$ is invariant with

respect to $\mathfrak{H}$, it follows that $t(\mathbf{hk}) = t(\mathbf{k})$ for each $\mathbf{h} \in \mathfrak{H}$ and $\mathbf{k} \in \mathfrak{T}$. Therefore, (20.4) equals

$$\max_{\mathbf{p}\in\mathscr{P}} \sum_{\mathbf{h}\in\mathfrak{H}} \sum_{\mathbf{k}\in\mathfrak{T}} \chi\{t(\mathbf{k}) \geq a\}\ p_{\mathbf{hk}} = \max_{\mathbf{p}\in\mathscr{P}} \sum_{\mathbf{k}\in\mathfrak{T}} \chi\{t(\mathbf{k}) \geq a\} \sum_{\mathbf{h}\in\mathfrak{H}} p_{\mathbf{hk}}$$
$$= \max_{\overline{\mathbf{p}}\in\mathscr{P}_{\mathfrak{T}}} \sum_{\mathbf{k}\in\mathfrak{T}} \chi\{t(\mathbf{k}) \geq a\}\ \overline{p}_{\mathbf{k}}.$$

Lemma 20.1 has various implications. The most basic is that biases, no matter how large, in picking $\mathbf{h} \in \mathfrak{H}$ have no impact at all on the sensitivity bound for test statistic $t(\mathbf{g})$ if $t(\cdot)$ is invariant with respect to $\mathfrak{H}$. In Tables 20.5 and 20.7, Wilcoxon's signed rank statistic is invariant to the permutations in $\mathfrak{H}$ in (20.2), so biases in how many cigarettes smokers smoke do not affect the sensitivity analysis for this statistic: the joint biases $\mathbf{p}$ in $\mathscr{P}$ only affect the smoker-control comparison through the corresponding marginal distributions $\overline{\mathbf{p}}$ in $\mathscr{P}_{\mathfrak{T}}$.

20.4.6 Symmetric Sensitivity Analyses for Biases That May Be Asymmetric

A symmetric set may contain objects that are individually asymmetric. Take the simplest example. Consider (x, y) as coordinates on the unit square, $[0, 1] \times [0, 1]$ or equivalently $\{(x, y) : 0 \leq x \leq 1,\ 0 \leq y \leq 1\}$. The unit square is a symmetric set: if (x, y) is in the unit square, then so is (y, x). On the unit square, the function $(x - y)^2$ is symmetric: it takes the same value at (x, y) as at (y, x). It sounds as if everything about $(x - y)^2$ on the unit square is symmetric, and in one sense that is true. The set of minima of $(x - y)^2$ on the unit square is $\{(x, x) : 0 \leq x \leq 1\}$, so this set contains elements (x, x) each of which is symmetric. On the other hand, the set of maxima of $(x - y)^2$ on the unit square is $\{(0, 1),\ (1, 0)\}$, so the set of maxima is symmetric, but its individual elements are asymmetric. Generally, symmetries are not all the same, and we need to be careful in connecting a symmetry in one place with a different symmetry in another. Specifically, a symmetric set may contain objects that are individually asymmetric.

In particular, if somewhat vaguely at first, a symmetric set of probability distributions $\mathscr{P}$ may contain distributions $\mathbf{p}$ that are decidedly asymmetric. To be a probability distribution on $\mathfrak{G}$, $\mathbf{p}$ must lie in a $|\mathfrak{G}| - 1$ dimensional simplex $\mathscr{S}$ of points with $p_{\mathbf{g}} \geq 0$ and $1 = \sum_{\mathbf{g}\in\mathfrak{G}} p_{\mathbf{g}}$. A sensitivity analysis may consider a family $\mathscr{P}$ of departures $\mathbf{p}$ from a completely randomized treatment assignment such that the family is symmetric, but the individual $\mathbf{p} \in \mathscr{P}$ are decidedly biased in one direction or another. For instance, in Tables 20.5 and 20.6, a small 7-dimensional sphere $\mathscr{P}$ interior to $\mathscr{S}$ centered at $\mathbf{p} = (1/8, \ldots, 1/8)^T$ has certain symmetries as a set, but only $(1/8, \ldots, 1/8)^T$ is symmetric in the sense of having equal coordinates. The situations considered so far are preliminary, slightly simpler

than the case that interests us, but these simpler situations are intended to guard against some incorrect intuitions.

Suppose that we had conducted two sensitivity analyses, perhaps one for a signed rank statistic applied to the smoker-minus-control pair differences, Y_i in Panel (i) of Fig. 19.2, and another for the crosscut statistic applied to the relationship between Y_i and the amount smoked, d_i, in Panel (ii) of Fig. 19.2. The first sensitivity concerns a marginal distribution of smoker/control assignments within pairs, $\mathbf{k} \in \mathfrak{T}$, while the second sensitivity analysis fixes $\mathbf{k}$, and concerns the assignment of doses, d_i, to matched pairs, or $\mathbf{h} \in \mathfrak{H}$. Expressed differently, the first sensitivity analysis found a bound on the P-value, say P, for a signed rank statistic as a maximum over a set $\mathscr{P}_{\mathfrak{T}}$ of possible marginal distributions $\overline{\mathbf{p}}$ for $\mathbf{k} \in \mathfrak{T}$, ignoring $\mathbf{h} \in \mathfrak{H}$. The second sensitivity analysis found a bound on P-value, say P', for a crosscut statistic as a maximum over a set $\mathscr{P}_{\mathfrak{H}}$ of possible conditional distributions of $\mathbf{h} \in \mathfrak{H}$, regarding $\mathbf{k} \in \mathfrak{T}$ as fixed. The task is to combine these two sensitivity analyses *without* introducing a new assumption about how $\mathbf{h} \in \mathfrak{H}$ and $\mathbf{k} \in \mathfrak{T}$ are related. Pointedly, the task is to avoid any assumption that $\mathbf{h} \in \mathfrak{H}$ and $\mathbf{k} \in \mathfrak{T}$ are unrelated or independent. The task is accomplished by forming the set of all joint distributions $\mathscr{P}$ on $\mathfrak{G} = \mathfrak{H}\mathfrak{T}$ that are compatible with the marginal distributions $\mathscr{P}_{\mathfrak{T}}$ for $\mathbf{k} \in \mathfrak{T}$ and with the conditional distributions $\mathscr{P}_{\mathfrak{H}}$ for $\mathbf{h} \in \mathfrak{H}$. Because $\mathscr{P}$ will contain every possible joint distribution, the restriction of $\mathbf{p}$ to $\mathscr{P}$ adds no assumption beyond the assumptions that the marginal distribution $\overline{\mathbf{p}}$ of $\mathbf{k}$ is in $\mathscr{P}_{\mathfrak{T}}$, and every conditional distributions $\widetilde{\mathbf{p}}_{\mathbf{k}}$ of $\mathbf{h}$ given $\mathbf{k}$ is in $\mathscr{P}_{\mathfrak{H}}$. In brief, $\mathscr{P}$ will be the set of all joint distributions compatible with what was already assumed in the two separate sensitivity analyses.

In greater detail, there is a subgroup $\mathfrak{H}$ of a group $\mathfrak{G}$ of $N \times N$ permutation matrices, and a transversal $\mathfrak{T} = \{\mathbf{k}_1, \ldots, \mathbf{k}_L\} \subseteq \mathfrak{G}$. The sensitivity analysis is concerned with a set $\mathscr{P}$ of distributions $\mathbf{p}$ on $\mathfrak{G}$. The set $\mathscr{P}$ is a kind of product, but not a direct product, of distributions on $\mathfrak{H}$ and $\mathfrak{T}$. A direct product would form $\mathscr{P}$ as composed from independent choices from $\mathfrak{H}$ and $\mathfrak{T}$, the product of a single distribution on $\mathfrak{H}$ and a single distribution on $\mathfrak{T}$, but that is not what we want. In Table 20.6, it seems entirely possible that the two biases are related: perhaps a person lacking self-discipline is more likely to be a daily smoker than a never-smoker, and also more likely to be a heavy smoker than a light smoker. There is no reason to think that $\mathbf{p} \in \mathscr{P}$ should make independent choices of $\mathbf{h} \in \mathfrak{H}$ and $\mathbf{k} \in \mathfrak{T}$ when picking $\mathbf{g} = \mathbf{hk} \in \mathfrak{G}$.

There is a set $\mathscr{P}_{\mathfrak{T}}$ of possible L-dimensional marginal distributions $\overline{\mathbf{p}}$ on $\mathfrak{T}$. Complete randomization on $\mathfrak{T}$ would entail that $\mathscr{P}_{\mathfrak{T}}$ contains a single element, namely the L-dimensional vector $\overline{\mathbf{p}} = (1/L, \ldots, 1/L)^T$, but larger sets $\mathscr{P}_{\mathfrak{T}}$ will permit us to entertain biased selection of $\mathbf{k}_\ell$ from $\mathfrak{T}$.

There is a second set $\mathscr{P}_{\mathfrak{H}}$ of $|\mathfrak{H}|$-dimensional conditional distributions $\widetilde{\mathbf{p}}_{\mathbf{k}}$ of $\mathbf{h} \in \mathfrak{H}$ given $\mathbf{k} \in \mathfrak{T}$. This second set $\mathscr{P}_{\mathfrak{H}}$ would entail complete randomization over $\mathfrak{H}$ given $\mathbf{k}$ if $\mathscr{P}_{\mathfrak{H}}$ contained only the single $|\mathfrak{H}|$-dimensional vector $\widetilde{\mathbf{p}}_{\mathbf{k}} = (1/|\mathfrak{H}|, \ldots, 1/|\mathfrak{H}|)^T$, but larger sets $\mathscr{P}_{\mathfrak{H}}$ will permit us to entertain biased selection of $\mathbf{h} \in \mathfrak{H}$ given $\mathbf{k} \in \mathfrak{T}$.

The set $\mathscr{P}$ of distributions on $\mathfrak{G}$ may now be described. Each $\mathbf{p} \in \mathscr{P}$ is formed by: (i) selecting one marginal distribution $\overline{\mathbf{p}} \in \mathscr{P}_{\mathfrak{T}}$ on $\mathfrak{T}$, and (ii) for each $\mathbf{k} \in \mathfrak{T}$ selecting a conditional distribution $\widetilde{\mathbf{p}}_{\mathbf{k}} \in \mathscr{P}_{\mathfrak{H}}$ of $\mathbf{h} \in \mathfrak{H}$ given $\mathbf{k} \in \mathfrak{T}$. By definition, the set $\mathscr{P}$ contains every distribution that can be formed in this way. Importantly, because we may, and often do, select a different $\widetilde{\mathbf{p}}_{\mathbf{k}} \in \mathscr{P}_{\mathfrak{H}}$ for different $\mathbf{k} \in \mathfrak{T}$, the distributions of $\mathbf{h} \in \mathfrak{H}$ and $\mathbf{k} \in \mathfrak{T}$ may be dependent, perhaps markedly so. The set $\mathscr{P}$ has certain symmetries: every $\overline{\mathbf{p}} \in \mathscr{P}_{\mathfrak{T}}$ is possible, and every $\widetilde{\mathbf{p}}_{\mathbf{k}} \in \mathscr{P}_{\mathfrak{H}}$ is possible for every $\mathbf{k} \in \mathfrak{T}$. However, the individual $\mathbf{p} \in \mathscr{P}$ may depart markedly from complete randomization with strong dependence between the components of $\mathbf{g} = \mathbf{hk}$. By analogy with the more algebraic discussion in Rosenbaum (2017), the set $\mathscr{P}$ of *all* distributions on $\mathfrak{G}$ that can be formed by (i) and (ii) above will be called the knit product of $\mathscr{P}_{\mathfrak{T}}$ and $\mathscr{P}_{\mathfrak{H}}$.[2]

20.4.7 Fixing One Evidence Factor

Suppose that we examined the two smoker-minus-control pair differences in periodontal disease in Table 20.7 and asked whether the larger difference occurred in the pair in which the smoker smoked 40 cigarettes per day in Table 20.5. More generally, we might ask whether the 441 smoker-minus-control differences in outcomes in Fig. 19.2 tend to be positively correlated with the amount smoked by the smokers in the various pairs, and in Sect. 19.4 we answered this question using the crosscut statistic. This analysis is fixing the identities of smoker and control within each pair, fixing $\mathbf{k} \in \mathfrak{T}$, and examining the arrangement of doses among the pairs, $\mathbf{h} \in \mathfrak{H}$. Lemma 20.2 gives a natural condition on the set $\mathscr{P}$ of joint distributions on the group $\mathfrak{G}$ such that casually regarding $\mathbf{k} \in \mathfrak{T}$ as fixed yields the same sensitivity bound as conditioning on the value of $\mathbf{k} \in \mathfrak{T}$. The condition is that $\mathscr{P}$ is a knit product of a set of conditional distributions $\mathscr{P}_{\mathfrak{H}}$ and a set of marginal distributions $\mathscr{P}_{\mathfrak{T}}$, so that the individual conditional distributions $\widetilde{\mathbf{p}}_{\mathbf{k}} \in \mathscr{P}_{\mathfrak{H}}$ do depend on $\mathbf{k}$ but the set $\mathscr{P}_{\mathfrak{H}}$ of these conditional distributions is the same set for all $\mathbf{k} \in \mathfrak{T}$. If $\widetilde{\mathbf{p}} \in \mathscr{P}_{\mathfrak{H}}$, then $\widetilde{\mathbf{p}} = \left(p_{\mathbf{h}_1}, \ldots, p_{\mathbf{h}_{|\mathscr{H}|}}\right)$ is one possible conditional distribution of $\mathbf{h}$ given $\mathbf{k}$. In Fig. 19.2, the statistic $t'(\mathbf{g})$ in Lemma 20.2 was the crosscut statistic.

[2]As just described, $\mathscr{P}$ is constructed from $\mathscr{P}_{\mathfrak{T}}$ and $\mathscr{P}_{\mathfrak{H}}$, which already exist. In this sense, the constructed $\mathscr{P}$ adds no assumption beyond what was already assumed by the two separate sensitivity analyses. An alternative starting point begins with an arbitrary set $\mathscr{Q}$ of joint distributions on $\mathfrak{G}$. This approach then continues by finding all of the possible marginal distribution of $\mathbf{k} \in \mathfrak{T}$ compatible with $\mathscr{Q}$. Continuing, this approach then finds the set $\mathscr{P}_{\mathfrak{H}}$ of all of the distributions $\widetilde{p}$ of $\mathbf{h} \in \mathfrak{H}$ that ever arise from $\mathscr{Q}$ as conditional distributions for $\mathbf{h}$ given $\mathbf{k}$. It concludes by constructing $\mathscr{P}$ as the knit product of $\mathscr{P}_{\mathfrak{T}}$ and $\mathscr{P}_{\mathfrak{H}}$. Clearly, $\mathscr{Q} \subseteq \mathscr{P}$. It follows that an upper bound on the two P-values, (P, P'), derived from $\mathscr{P}$ is also an upper bound on the two P-values, (P, P'), derived from $\mathscr{Q}$. In practice, we begin with two sensitivity analyses, one for each factor separately, and we seek to combine them.

Lemma 20.2 *Suppose a set $\mathscr{P}$ of distributions on group $\mathfrak{G}$ is the knit product of conditional distributions $\mathscr{P}_{\mathfrak{H}}$ on a subgroup $\mathfrak{H}$ and marginal distributions $\mathscr{P}_{\mathfrak{T}}$ on a transversal $\mathfrak{T}$. Then the conditional sensitivity bound from $\mathscr{P}$ for a statistic $t'(\mathbf{g}) = t'(\mathbf{hk})$ given $\mathbf{k}$, namely*

$$\max_{\mathbf{p}\in\mathscr{P}} \frac{\sum_{\mathbf{h}\in\mathfrak{H}} \chi\left\{t'(\mathbf{hk}) \geq a\right\} p_{\mathbf{hk}}}{\sum_{\mathbf{h}\in\mathfrak{H}} p_{\mathbf{hk}}}, \tag{20.6}$$

equals the sensitivity bound from $\mathscr{P}_{\mathfrak{H}}$ that simply views $\mathbf{k}$ as fixed, namely

$$\max_{\widetilde{\mathbf{p}}\in\mathscr{P}_{\mathfrak{H}}} \sum_{\mathbf{h}\in\mathfrak{H}} \chi\left\{t'(\mathbf{hk}) \geq a\right\} \widetilde{p}_{\mathbf{h}}. \tag{20.7}$$

Proof Because $\mathscr{P}$ is the knit product of the conditional distributions $\left(p_{\mathbf{h}_1}, \ldots, p_{\mathbf{h}_{|\mathscr{H}|}}\right) \in \mathscr{P}_{\mathfrak{H}}$ and marginal distributions $\overline{\mathbf{p}} \in \mathscr{P}_{\mathfrak{T}}$, the set of possible conditional distributions $p_{\mathbf{hk}}/\sum_{\mathbf{h}\in\mathfrak{H}} p_{\mathbf{hk}}$ of $\mathbf{h} \in \mathfrak{H}$ given $\mathbf{k}$ is precisely $\mathscr{P}_{\mathfrak{H}}$ for every $\mathbf{k} \in \mathfrak{T}$. Hence, no matter which $\mathbf{k}$ occurs, the set of probability distributions considered in (20.6) is the same as the set of distributions in (20.7), and the maxima are equal.

20.4.8 Joint Behavior of Two Bounds on P-Values

The following elementary lemma is concerned with the joint behavior of a bound on a P-value and a bound on a conditional P-value. It is a special case of a more general but equally elementary lemma in [31]. To be specific, in Fig. 19.2, the statistic $t(\mathbf{g})$ might be Wilcoxon's signed rank statistic yielding the P-value P in Lemma 20.1, and the statistic $t'(\mathbf{g})$ might be the crosscut statistic yielding the P-value P' in Lemma 20.2.

Lemma 20.3 *Suppose that P is a function of $\mathbf{K}$ and P' is a function of $(\mathbf{H}, \mathbf{K})$ such that $\Pr(P \leq \alpha) \leq \alpha$ for all $0 \leq \alpha \leq 1$ and $\Pr\left(P' \leq \alpha' \mid \mathbf{K} = \mathbf{k}\right) \leq \alpha'$ for each $\mathbf{k}$ and for all $0 \leq \alpha' \leq 1$. Then $\left(P, P'\right)$ is stochastically larger than the uniform distribution on the unit square.*

Proof Let $\left(Q, Q'\right)$ be uniform on the unit square, hence independent, so the task is to show that $\left(P, P'\right)$ is stochastically larger than $\left(Q, Q'\right)$, where $\Pr\left(Q \leq \alpha,\ Q' \leq \alpha'\right) = \alpha\alpha'$. By assumption, $\Pr(P \leq \alpha) \leq \alpha = \Pr(Q \leq \alpha)$. Using $\Pr\left(P' \leq \alpha' \mid \mathbf{K}\right) \leq \alpha'$, the independence of Q and Q', and the fact that P is a function of $\mathbf{K}$, we have

$$\Pr\left(P' \leq \alpha' \mid P\right) = \mathrm{E}\left\{\Pr\left(P' \leq \alpha' \mid P, \mathbf{K}\right) \mid P\right\}$$

$$= \mathrm{E}\left\{\Pr\left(P' \leq \alpha' \mid \mathbf{K}\right) \mid P\right\} \leq \alpha' = \Pr\left(Q' \leq \alpha' \mid Q\right).$$

The result follows from Cohen and Sackrowitz [4, Theorem 2.5] and the independence of Q and Q'.

In the context of adaptive design of clinical trials, Brannath, Posch, and Bauer [2] discuss bivariate P-values, $\left(P, P'\right)$, that are stochastically larger than the uniform distribution on the unit square. Notice that the conclusion that $\Pr\left(P \leq \alpha,\ P' \leq \alpha'\right) \leq \alpha\alpha'$ for all $0 \leq \alpha \leq 1$ and $0 \leq \alpha' \leq 1$ does not imply that $\left(P, P'\right)$ are independent; rather, it implies that $\left(P, P'\right)$ have a possibly dependent joint distribution that is larger than the joint distribution of two independent uniform random variables.

20.4.9 Joint Distribution of P-Value Bounds for Two Evidence Factors

We observe a particular treatment assignment $\mathbf{G} \in \mathfrak{G}$ having an unknown distribution $\mathbf{p}$ in a known set $\mathscr{P}$. Were Fisher's hypothesis H_0 of no effect true, then $\mathbf{g}$ would have rearranged the N experimental subjects, assigning them to roles in the study design, but the assignment of individuals to treatments would not alter their outcomes. The group $\mathfrak{G}$ has a subgroup $\mathfrak{H}$ whose cosets have a transversal $\mathfrak{T} = \{\mathbf{k}_1, \ldots, \mathbf{k}_L\} \subseteq \mathfrak{G}$, so each $\mathbf{g} \in \mathfrak{G}$ has a unique representation as $\mathbf{g} = \mathbf{hk}$ with $\mathbf{h} \in \mathfrak{H}$ and $\mathbf{k} \in \mathfrak{T}$. We realize a particular $\mathbf{G} = \mathbf{HK}$, selected according to the unknown distribution $\mathbf{p}$, yielding the observed values of two test statistics, $t(\mathbf{G})$ and $t'(\mathbf{G})$, where $t(\cdot)$ is $\mathfrak{H}$-invariant, so $t(\mathbf{hk}) = t(\mathbf{k})$. For instance, in Fig. 19.2, the statistic $t(\mathbf{g})$ might be Wilcoxon's signed rank statistic, and the statistic $t'(\mathbf{g})$ might be the crosscut statistic.

Were $\mathbf{p}$ known, we could compute the two corresponding true P-values, the marginal P-value from $t(\mathbf{G})$,

$$P = \sum_{\mathbf{g} \in \mathfrak{G}} \chi\left\{t(\mathbf{g}) \geq a\right\} p_{\mathbf{g}} \text{ with } a = t(\mathbf{G}),$$

and the conditional P-value given $\mathbf{K}$ from $t'(\mathbf{G})$,

$$P' = \frac{\sum_{\mathbf{h} \in \mathfrak{H}} \chi\left\{t'(\mathbf{hk}) \geq a'\right\} p_{\mathbf{hk}}}{\sum_{\mathbf{h} \in \mathfrak{H}} p_{\mathbf{hk}}} \text{ with } a' = t'(\mathbf{G}) \text{ and } \mathbf{k} = \mathbf{K}.$$

Because $\mathbf{p}$ is not known, we compute two separate sensitivity bounds, (20.5) with $a = t(\mathbf{G})$, namely

$$\overline{P} = \max_{\overline{\mathbf{p}} \in \mathscr{P}_{\mathfrak{T}}} \sum_{\mathbf{k} \in \mathfrak{T}} \chi\left\{t(\mathbf{k}) \geq t(\mathbf{K})\right\} \overline{p}_{\mathbf{k}},$$

and (20.7) with $a = t'(\mathbf{G})$ and $\mathbf{k} = \mathbf{K}$, namely

$$\overline{P}' = \max_{\widetilde{\mathbf{p}} \in \mathscr{P}_{\mathfrak{H}}} \sum_{\mathbf{h} \in \mathfrak{H}} \chi \left\{ t'(\mathbf{hK}) \geq t'(\mathbf{HK}) \right\} \widetilde{p}_{\mathbf{h}}.$$

Proposition 20.2 *Assume Fisher's null hypothesis H_0 of no treatment effect is true for the purpose of testing it. Let $\mathfrak{G}$ be a group of $N \times N$ permutation matrices, let $\mathfrak{H}$ be a subgroup of $\mathfrak{G}$, and let $\mathfrak{T} = \{\mathbf{k}_1, \ldots, \mathbf{k}_L\} \subseteq \mathfrak{G}$ be a transversal of the cosets of $\mathfrak{H}$. Let $\mathscr{P}_{\mathfrak{H}}$ be a closed set of probability distributions on $\mathfrak{H}$, let $\mathscr{P}_{\mathfrak{T}}$ be a closed set of probability distributions on $\mathfrak{T}$, and let $\mathscr{P}$ be the knit product of $\mathscr{P}_{\mathfrak{H}}$ and $\mathscr{P}_{\mathfrak{T}}$, so $\mathscr{P}$ is a set of distributions on $\mathfrak{G}$. Sample $\mathbf{G} = \mathbf{HK} \in \mathfrak{G}$ with some unknown distribution $\mathbf{p} \in \mathscr{P}$. Let $t(\cdot)$ and $t'(\cdot)$ be two test statistics, that is, two real-valued functions defined on $\mathfrak{G}$, where $t(\cdot)$ is $\mathfrak{H}$-invariant. Then $\left(\overline{P}, \overline{P}'\right)$ is stochastically larger than the uniform distribution on the unit square.*

Proof The unknown but true P-values $\left(P, P'\right)$ satisfy the conditions of Lemma 20.3, so $\left(P, P'\right)$ is stochastically larger than the uniform distribution on the unit square. By Lemma 20.1, $\overline{P} \geq P$. By Lemma 20.2, $\overline{P}' \geq P'$. So $\left(\overline{P}, \overline{P}'\right)$ is stochastically larger than the uniform distribution on the unit square.

20.4.10 Another Example: Blocks of Size Three

So far, in this section, the example has been the tiny version of Fig. 19.2 in Tables 20.5, 20.6, 20.7. For this tiny example and for Fig. 19.2, Proposition 20.2 yields the conclusion that a signed rank statistic and a crosscut statistic constitute two evidence factors. What does Proposition 20.2 say about the tannery data in Table 20.1?

Consider one block of three individuals in Table 20.1, perhaps the first block. The three individuals in this block may be assigned to treatments E1, E2, and C in 3!=6 ways, and these 6 assignments correspond with the symmetric group $\mathfrak{G}$ of all $|\mathfrak{G}| = 6$ permutation matrices of size 3×3. One subgroup of $\mathfrak{G}$ is $\mathfrak{H}$ given by:

$$\mathfrak{H} = \left\{ \mathbf{h}_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \ \mathbf{h}_2 = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \right\}. \tag{20.8}$$

Recall that $\mathbf{n}$ is the 3×1 column vector $\mathbf{n} = (1, 2, 3)^T$. In (20.8), the second permutation matrix, $\mathbf{h}_2$, swaps the two individuals assigned to E1 and E2. For instance, $\mathbf{h}_2\mathbf{n} = (2, 1, 3)^T$.

Now define

$$\mathfrak{K} = \left\{ \mathbf{k}_1 = \begin{bmatrix} 1\,0\,0 \\ 0\,1\,0 \\ 0\,0\,1 \end{bmatrix},\ \mathbf{k}_2 = \begin{bmatrix} 0\,0\,1 \\ 1\,0\,0 \\ 0\,1\,0 \end{bmatrix},\ \mathbf{k}_3 = \begin{bmatrix} 0\,1\,0 \\ 0\,0\,1 \\ 1\,0\,0 \end{bmatrix} \right\}. \tag{20.9}$$

In (20.8), $\mathbf{k}_1\mathbf{n} = (1, 2, 3)^T$ places unit 3 in C, $\mathbf{k}_2\mathbf{n} = (3, 1, 2)^T$ places unit 2 in C, and $\mathbf{k}_3\mathbf{n} = (2, 3, 1)^T$ places unit 1 in C. Each of the six $\mathbf{g} \in \mathfrak{G}$ has a unique representation as $\mathbf{g} = \mathbf{hk}$ with $\mathbf{h} \in \mathfrak{H}$ and $\mathbf{k} \in \mathfrak{K}$, and $\mathfrak{K}$ is a transversal of the cosets of $\mathfrak{H}$. In words, $\mathfrak{K}$ picks one person for treatment C, then $\mathfrak{H}$ assigns the two remaining people to E1 and E2.

Consider the stratified Wilcoxon statistic comparing group C to $\{E1, E2, C\}$. That statistic is affected by the identity of the individual assigned to C, but once that is settled, the statistic ignores who is assigned to E1 and who is assigned to E2. Although not a rank statistic, the permutational m-statistic comparing group C to $\{E1, E2, C\}$ behaves in the same way. Generally, a statistic $t(\mathbf{g}) = t(\mathbf{hk})$ that compares C to $\{E1, E2, C\}$, ignoring who is assigned to E1 versus E2 is $\mathfrak{H}$-invariant; that is, $t(\mathbf{hk}) = t(\mathbf{k})$ for each $\mathbf{h} \in \mathfrak{H}$ and $\mathbf{k} \in \mathfrak{K}$.

In (20.8) and (20.9), if we sample $\mathbf{G} = \mathbf{HK} \in \mathfrak{G}$, and then condition on the value of $\mathbf{K}$, then we have fixed the identity of the individual in treatment C, leaving $\mathbf{H}$ to swap the two remaining individuals among E1 and E2. Given $\mathbf{K}$, the permutation distribution of a statistic $t'(\mathbf{G})$ is examining the effects of changing the treatment assignments for the two individuals not in treatment group C. Here, $t'(\mathbf{G})$ might be the stratified Wilcoxon statistic or the m-statistic comparing E2 to $\{E1, E2\}$, holding fixed the identity of the third individual in C.

So far, we have considered the meaning of Proposition 20.2 for one block in Table 20.1, but there are 30 blocks in this table. A randomized block experiment would make 30 independent treatment assignments in the 30 blocks, each assignment having the structure above. Each such assignment is a 90×90 permutation matrix, where each matrix is block diagonal, with 30 blocks each consisting of a 3×3 permutation matrix. The relevant groups, $\mathfrak{G}$, $\mathfrak{H}$, and $\mathfrak{K}$ are the 30-fold direct products of (20.8) and (20.9). For instance, $\mathfrak{G}$ contains $(3!)^{30}$ block diagonal 90×90 permutation matrices with 30 blocks of size 3×3. Similarly, $\mathfrak{H}$ contains 2^{30} such matrices, $\mathfrak{K}$ contains 3^{30}, and $\mathfrak{G} = \mathfrak{H}\mathfrak{K}$.

20.4.11 Untidy Blocks

The design in Table 20.1 is tidy: every block is the same size and every block assigns one individual to E1, E2, and C. Tidy blocks are not central to the argument. One block could have one E1, two E2's, and one C, the next block could have five E1's, no E2, and three C's, and so on. In this case, the permutation group that acts on the first block is different from the permutation group that acts on the second, so

the block diagonal permutation matrix that acts on the entire design has blocks of different sizes and structures.

20.5 Further Reading

For the basic theory of evidence factors, see [30, 31, 36]. For another example, see [15]. Bikram Karmakar and colleagues [13] discuss the Bahadur efficiency of the combination of two sensitivity analyses formed from two evidence factors. Use of the computer to build a matched design with evidence factors is discussed in Karmakar et al. [14]. Zubizarreta and colleagues [46] build two evidence factors in a different way: one factor compares two treatments within institutions – specifically, within hospitals – that use both treatments; the second factor compares institutions that almost invariably use one of the two treatments. The selection biases that operate within an institution may be very different from the selection biases that guide an individual to an institution, so these two factors are likely to be affected by different biases. Kai Zhang and colleagues [45] consider two evidence factors, one defined by time, the other by geography: some babies were spared the treatment because of when they were born, others were spared because of where they were born.

20.6 Software

An analysis using evidence factors combines two sensitivity analyses that are essentially independent. Once one has demonstrated that the two analyses are essentially independent in the sense of Proposition 20.2, these two separate analyses are done using conventional sensitivity analyses for permutation inferences. Only the combination step requires additional software. For an analysis using `R` to conduct an analysis using two evidence factors, see [34, §4].

The truncated product of P-values of [43] is implemented as either `truncatedP` or `truncatedPbg` in the `R` package `sensitivitymv`. The two functions are two formulas for the same calculation, so they return the same results. The function `truncatedP` uses the formula in [43], while the function `truncatedPbg` uses the binomial mixture of gamma distributions in [10, Expression (6)]. As a special case, these functions implement Fisher's method of combining P-values.

20.7 Data

The tannery data of [44] are contained in the `R` package `DOS2` as `tannery`, and its analysis is given in the help-file for `tannery`.

The periodontal disease and smoking data are contained in the `R` package `DOS2` as `periodontal` and its analysis appears in the help-file for the `crosscut` function in that package.

Bibliography

1. Alam, K.: Some nonparametric tests of randomness. J. Am. Stat. Assoc. **69**, 738–739 (1974)
2. Brannath, W., Posch, M., Bauer, P.: Recursive combination tests. J. Am. Stat. Assoc. **97** 236–244 (2002)
3. Cochran, W.G.: The planning of observational studies of human populations (with Discussion). J. R. Stat. Soc. A **128**, 234–265 (1965)
4. Cohen, A., Sackrowitz, H.B.: On stochastic ordering of random vectors. J. Appl. Prob. **32**, 960–965 (1995)
5. Collins, A.R: The comet assay for DNA damage and repair: principles, applications, and limitations. Mol. Biotech. **26**, 249–261 (2004)
6. Conover, W.J., Salsburg, D.S.: Locally most powerful tests for detecting treatment effects when only a subset of patients can be expected to 'respond' to treatment. Biometrics **44**, 189–196 (1988)
7. Dwass, M.: Some k-sample rank-order tests. In: Contributions to Probability and Statistics, pp. 198–202. Stanford University Press, Stanford, CA (1960)
8. Hodges, J.L., Lehmann, E.L.: The efficiency of some nonparametric competitors of the t-test. Ann. Math. Stat. **27**, 324–335 (1956)
9. Hollander, M., Pledger, G., Lin, P.E.: Robustness of the Wilcoxon test to a certain dependency between samples. Ann. Stat. **2**, 177–181 (1974)
10. Hsu, J.Y., Small, D.S., Rosenbaum, P.R.: Effect modification and design sensitivity in observational studies. J. Am. Stat. Assoc. **108**, 135–148 (2013)
11. Huber, P.J.: Robust estimation of a location parameter. Ann. Math. Stat. **35**, 73–101 (1964)
12. Isaacs, I.M.: Algebra: A Graduate Course. American Mathematical Society, Providence, RI (1994)
13. Karmakar, B., French, B., Small, D.S.: Integrating the evidence from evidence factors in observational studies. Biometrika **106**, 353–367 (2019)
14. Karmakar, B., Small, D.S., Rosenbaum, P.R.: Using approximation algorithms to build evidence factors and related designs for observational studies. J. Comput. Graph. Stat. **28**, 698–709 (2019)
15. Karmakar, B., Small, D.S., Rosenbaum, P.R.: Using evidence factors to clarify exposure biomarkers. Am. J. Epidemiol. **189**, 243–249 (2020)
16. Kopjar, N, Garaj-Vrhovac, V.: Application of the alkaline comet assay in human biomonitoring for genotoxicity: a study on Croatian medical personnel handling antineoplastic drugs. Mutagenesis **16**, 71–78 (2001)
17. Kurzweil, H., Stellmacher, B.: The Theory of Finite Groups: An Introduction. Springer, New York (2004)
18. Lehmann, E.L.: Nonparametrics: Statistical Methods Based on Ranks. Holden-Day, San Francisco (1975)
19. Marcus, R., Peritz, E., Gabriel, K.R.: On closed testing procedures with special reference to ordered analysis of variance. Biometrika **63**, 655–60 (1976)

20. Marden, J.I.: Use of nested orthogonal contrasts in analyzing rank data. J. Am. Stat. Assoc. **87**, 307–318 (1992)
21. Maritz, J.S.: A note on exact robust confidence intervals for location. Biometrika **66**, 163–166 (1979)
22. Maritz, J.S.: Distribution-Free Statistical Methods. Chapman and Hall, London (1995)
23. Olive, P.L., Banath, J.P.: The comet assay: a method to measure DNA damage in individual cells. Nat. Protoc. **1**, 23–29 (2006)
24. Peirce, C.S.: Some consequences of four incapacities. J. Specul. Philos. **2**, 140–157 (1868). Reprinted in 2011 in: Talisse, R.B., Aikin, S.F. (eds.) The Pragmatism Reader: From Peirce through the Present. Harvard University Press, Cambridge, MA
25. Randles, R.H., Hogg, R.V.: Certain uncorrelated statistics and independent rank statistics. J. Am. Stat. Assoc. **66**, 569–574 (1971)
26. Resnick, S.I.: A Probability Path. Springer, New York (2014)
27. Rosenbaum, P.R.: Replicating effects and biases. Am. Stat. **55**, 223–227 (2001)
28. Rosenbaum, P.R.: Sensitivity analysis for m-estimates, tests, and confidence intervals in matched observational studies. Biometrics **63**, 456–464 (2007)
29. Rosenbaum, P.R.: Confidence intervals for uncommon but dramatic responses to treatment. Biometrics **63**, 1164–1171 (2007)
30. Rosenbaum, P.R.: Evidence factors in observational studies. Biometrika **97**, 333–345 (2010)
31. Rosenbaum, P.R.: Some approximate evidence factors in observational studies. J. Am. Stat. Assoc. **106**, 285–295 (2011)
32. Rosenbaum, P.R.: A new U-statistic with superior design sensitivity in matched observational studies. Biometrics **67**, 1017–1027 (2011)
33. Rosenbaum, P.R.: An exact adaptive test with superior design sensitivity in an observational study of treatments for ovarian cancer. Ann. Appl. Stat. **6**, 83–105 (2012)
34. Rosenbaum, P.R.: Two R packages for sensitivity analysis in observational studies. Observational Stud. **1**, 1–17 (2015)
35. Rosenbaum. P.R.: How to see more in observational studies: Some new quasi-experimental devices. Ann. Rev. Stat. Appl. **2**, 21–48 (2015)
36. Rosenbaum, P.R.: The general structure of evidence factors in observational studies. Stat. Sci. **32**, 514–530 (2017)
37. Rosenbaum, P.R., Small, D.S.: An adaptive Mantel–Haenszel test for sensitivity analysis in observational studies. Biometrics **73**, 422–430 (2017)
38. Rosenbaum, P.R.: Sensitivity analysis for stratified comparisons in an observational study of the effect of smoking on homocysteine levels. Ann. Appl. Stat. **12**, 2312–2334 (2018)
39. Susser, M.: Causal Thinking in the Health Sciences: Concepts and Strategies in Epidemiology. Oxford University Press, New York (1973)
40. Susser, M.: Falsification, verification and causal inference in epidemiology: Reconsideration in the light of Sir Karl Popper's philosophy. In: Susser, M. (ed.), Epidemiology, Health and Society: Selected Papers, pp. 82–93. Oxford University Press, New York (1987)
41. Wilcoxon, F.: Individual comparisons by ranking methods. Biometrics **1**, 80–83 (1945)
42. Wolfe, D.A.: Some general results about uncorrelated statistics. J. Am. Stat. Assoc. **68**, 1013–1018 (1973)
43. Zaykin, D.V., Zhivotovsky, L.A., Westfall, P.H., Weir, B.S.: Truncated product method for combining P-values. Genet. Epidemiol. **22**, 170–185 (2002)
44. Zhang, M., Chen, Z., Chen, Q., Zou, H., Lou, J. He, J.: Investigating DNA damage in tannery workers occupationally exposed to trivalent chromium using comet assay. Mutat. Res. Genet. Toxicol. Environ. Mutagen. **654**, 45–51 (2008)
45. Zhang, K., Small, D.S., Lorch, S., Srinivas, S., Rosenbaum, P.R.: Using split samples and evidence factors in an observational study of neonatal outcomes. J. Am. Stat. Assoc. **106**, 511–524 (2011)
46. Zubizarreta, J.R., Neuman, M.D., Silber, J. H., Rosenbaum, P.R.: Contrasting evidence within and between institutions that provide treatment in an observational study of alternate forms of anesthesia. J. Am. Stat. Assoc. **107**, 901–915 (2012)

Chapter 21
Constructing Several Comparison Groups

Abstract Evidence that a treatment is actually the cause of its ostensible effects is often strengthened by comparisons involving several groups. Can the computer help in constructing these groups? Several methods are described: the exterior match, optimal tapered matching, approximation algorithms for matching several groups, and the possibility of attenuating bias by leaving some covariates unmatched.

21.1 Why Compare Several Groups?

21.1.1 Uses of Several Comparison Groups

The use of more than one comparison group has been illustrated several times. For instance, Sects. 5.2.2 and 12.3 discussed the possibility of detecting and appraising unmeasured bias using two or more control groups [4, 10, 13–15, 24]. In [18], a direct comparison of treated and control groups was strengthened by adding a comparison of the differential effects of two treatments promoted by the same generic unobserved bias; see Sect. 5.2.6 and [16]. Chapter 20 used comparisons among several groups as evidence factors.

Tapered matching uses multiple groups to explore the consequences of deliberately leaving an observed covariate unmatched [6, 21]. For instance, Rachel Kelz et al. [8] compared the health outcomes of patients receiving surgery performed by newly trained surgeons or by experienced surgeons working at the same hospitals in the USA. Newly trained surgeons had recently completed their surgical training. Outcomes, including mortality, were much worse for patients of newly trained surgeons when the match controlled only for a few covariates such as the surgical procedure; however, the difference in mortality was negligible when the match controlled for the patients' other health problems and for emergency admission. Newly trained surgeons often treated some of the sickest and most complex patients, often in emergencies at off-hours, and experienced surgeons did negligibly better with these same patients, an interesting fact brought to light tapered matching.

It is sometimes suggested that leaving an unimportant covariate unmatched will attenuate unmeasured bias [3, 20], a possibility discussed in Sect. 21.5 and [11].

© Springer Nature Switzerland AG 2020

P. R. Rosenbaum, *Design of Observational Studies*, Springer Series in Statistics,
https://doi.org/10.1007/978-3-030-46405-9_21

Can the computer be used to construct comparisons with more than one group? Nonbipartite matching was used in Sect. 12.3 to form matched pairs from three groups in a balanced incomplete block design [9]. In such an incomplete block design, each pair contains two treatments, not three treatments, but any two treatments do occur together in some pairs. Are their other approaches besides incomplete block designs?

21.1.2 Issues When Constructing More Than One Comparison Group

A simple approach, discussed in Sect. 21.2, builds two comparison groups one at a time, allowing the two groups to overlap, so some controls appear twice, once in each group. This simple approach has both advantages and disadvantages. One advantage is that the closest controls are available for use in both groups, an important consideration when close controls are in short supply. A second advantage is that the construction of each comparison group does not alter the construction of the other comparison group, an important consideration when the two comparison groups are each intended to represent some natural population. A disadvantage is that comparisons of the two control groups must deal with the overlap. A device for dealing with this disadvantage is the exterior match [19] in Sect. 21.2.

The three-group minimum distance incomplete block design in Sect. 12.3 may be constructed quickly, in polynomial time [9]; however, it produces a balanced set of matched pairs, not triples. Minimizing the total distance when forming complete blocks—matched triples—is computationally difficult; specifically, the problem is NP-complete so it is believed that no polynomial time algorithm can solve it [5]. An approximation algorithm produces in polynomial time a solution to an optimization problem such that the solution cannot be much worse than the unobtainable optimal solution [25, 27]. An approximation algorithm for matched triples is discussed in Sect. 21.4 and [5, 7].

21.2 Overlapping Comparison Groups and the Exterior Match

21.2.1 Two Entwined Comparison Groups

The simplest way to build two matched comparison groups is to match each treated individual to one control by a certain standard, then to repeat the process with the same treated individuals but possibly different matched controls selected by a different standard. That is, controls are twice selected, by different standards, from a single reservoir of potential controls. This process implicitly produces

matched triples anchored by the same treated individual. For instance, as discussed in Sect. 21.1, in a tapered match, Kelz et al. [8] matched the patients of newly trained surgeons to the patients of experienced surgeons, controlling for many more covariates in second of the two matches. An important conclusion reached in [8] was that the higher mortality among the patients of newly trained surgeons reflected the much poorer initial health and emergent status of their patients, a conclusion reached by comparing two comparison groups of patients treated by experienced surgeons. For other examples of tapered matching in disparities research, see [21–23].

Because the investigator dips twice into the same reservoir of potential controls, some of the same controls may show up in both control groups. Two control groups that are matched in pairs to the same treated group are said to be entwined if some of the same controls appear in both groups. This presents no problems when the control groups are used one at a time. However, we often wish to ask a basic question: Do the two control groups differ in their outcomes? For instance, this was an important question in [8, 21]. To answer this basic question, the overlap in the two control groups must be addressed. This is done using the exterior match [19].

21.2.2 The Exterior Match

To summarize, one treated group has been matched twice, yielding two controls for the first treated individual, two controls for the second treated individual, and so on. So, there are matched triples anchored by the same treated individual. Hence, the two controls from the two control groups are paired with each other by virtue of being paired with the same treated individual. One person may show up twice, once as a control in the first group, a second time as a control in the second group. Perhaps both copies of that one duplicated person are paired with each other; however, often this is not the case, and the two copies are paired with different treated individuals. In [8, 21], the two control groups were formed by matching for different sets of covariates, so duplicated controls were often matched to different treated individuals.

Here are a few obvious considerations concerning the basic question about whether outcomes differ in the two control groups. If the two standards that defined the two control groups picked out exactly the same controls, then of course the outcomes in the two control groups do not differ. The more the control groups overlap, the less opportunity there is for the control groups to exhibit different outcomes. If matching for income picks the same controls as matching for income and education, then the control group matched for income alone has the same outcomes as the control group matched for both income and education. So, the degree of overlap is part of the answer to the basic question. The other part of the answer concerns the nonoverlapping portion, the controls who appear in exactly one control group: Are their outcomes similar or different? It makes no sense to apply a statistical test to the two control groups in their entirety, because that would be asking whether a person in the overlap differs from herself; of course, she does

not. We could simply remove from both groups individuals who appear in both control groups, but in general this breaks the matched pairs, leaving behind the broken remnants of matched data set. The exterior match [19] solves the problem by returning a new matched sample, without the overlap, with the pairing of individuals adjusted as little as possible.

Suppose that there are I treated individuals, hence I matched triples, hence I pairs i of two controls from the two control groups, $i = 1, \ldots, I$. Let $a_1, \ldots, a_I$ be the unique identification numbers—say, the Social Security Numbers—of the I controls in the first control group, and let $b_1, \ldots, b_I$ be the unique identification numbers of the I controls in the second control group. Here, person a_i is paired with person b_i for $i = 1, \ldots, I$. If $a_i = b_j$ for some i and j, then person $a_i = b_j$ is one of the controls who appear in both control groups.

The exterior match removes the overlap while minimally adjusting the pairing. How is this done? If both a_i and b_i appear in only one control group, then we retain pair (a_i, b_i) for the exterior match. If $a_i = b_i$, then we remove pair (a_i, b_i), so it is not part of the exterior match. Those are the obvious cases. Now consider the remaining cases.

If a_i appears only in control group one, but b_i appears in both control groups, then we want to retain a_i but remove b_i. So, $b_i = a_j$ for some $j \neq i$. If b_j appears only in the second control group, then we remove b_i and form a new matched pair (a_i, b_j). What if b_j also appears in both groups? Then, $b_j = a_k$ for some $k \neq j$, $k \neq i$. If b_k appears only in the second control group, then we remove b_i and b_j, and form a new matched pair (a_i, b_k). This process continues, but must ultimately produce some b_n who appears only in the second control group, yielding a new matched pair (a_i, b_n).

The process in the previous paragraph is repeated for each a_i who appears only in control group one. The resulting pairs are the exterior match. When this process is complete, there may be some remaining a_i who appears in both groups. If we applied the steps in the previous paragraph to this a_i, the process would eventually come back to a_i forming a cycle of duplicated individuals, all of whom are removed.

The process just described is performed by the `exteriorMatch` package in `R`. Specifically, `exteriorMatch` takes as input the two columns, $a_1, \ldots, a_I$ and $b_1, \ldots, b_I$, and it returns the exterior match.

Are outcomes similar in the two control groups? This basic question is answered by a matched-pair analysis of the exterior match, noting also the extent to which duplicate controls were removed.

The exterior match preserves some of the properties of the original match. If individuals were exactly matched for gender, then the exterior match is exactly matched for gender. If the original match finely balanced a ten-category nominal variable, say deciles of the propensity score, then the exterior match finely balances this same nominal variable. For these and other properties of the exterior match, see [19].

21.3 Optimal Tapered Matching

The simple method in Sect. 21.2 formed two comparison groups, one at a time, permitting the two comparison groups to overlap. The method is easy to implement and it extends immediately to three or more comparison groups. This method prevents the construction of one comparison group from affecting the construction of another. These overlapping comparison groups may be compared to each other using the exterior match.

When the pool of potential controls is large, an alternative method builds two comparison groups at the same time, preventing overlap [6]. The alternative method optimally assigns potential controls to one comparison group or the other, while also optimally pairing each treated individual to two controls, one from each group.

This alternative method has advantages and disadvantages when compared to the method in Sect. 21.2. By removing duplicates in Sect. 21.2, the exterior match is smaller than the two comparisons with the treated group, so the comparison of the two control groups will typically have lower power than each of the treatment-control comparisons. In contrast, preventing overlap produces three nonoverlapping groups of the same size. Moreover, removing duplicates means that the exterior match may describe a different population than the treated group. For instance, if the treated group is mostly female and the control reservoir is mostly male, then the method in Sect. 21.2 is more likely to use the same female in the two comparison groups than to use the same male; so, females are more likely than males to be removed as duplicates when forming the exterior match.

The main disadvantage of preventing overlap is that each comparison group takes away controls who might be useful in the other comparison group. This depletion may take away some of the closest controls, and the presence of the second control group may alter the population represented by the first control group.

Optimal tapered matching is described in detail in the work of Shoshana Daniel et al. [6]. A brief summary follows, but additional specifics in the paper should be considered before applying the method. Suppose that there are T treated individuals and $C \geq 2 \times T$ potential controls available in the control reservoir. We want to match each treated individual to two controls to produce two comparison groups defined by different criteria.

There are two criteria for matching that define the two potential comparison groups. Each criterion is defined by a $T \times C$ distance matrix. For instance, one distance matrix might be a Mahalanobis distance based on a few covariates, and the other distance matrix might be a Mahalanobis distance matrix based on those covariates plus additional covariates. Often the distance matrices are penalized for almost-exact matching or for calipers on scores. See [6] for discussion of "compatible" distance matrices.

The two $T \times C$ distance matrices are stacked, one on top of the other, making one $2T \times C$ distance matrix. For this $2T \times C$ distance matrix, solve the optimal assignment problem, pairing each row with a different column to minimize the total distance. This can be done using the `pairmatch` function in the `optmatch`

package in R. It is at this step that the two distance matrices must be "compatible," because the optimal assignment algorithm is minimizing the sum of $2T$ distances, T from the first distance matrix, T more from the second distance matrix. So, "compatible" means that this total is a reasonable criterion to minimize, as indeed it is when the two distance matrices are two Mahalanobis distance matrices based on an enlarged set of covariates; see [6, §3.4].

An optimal assignment for the $2T \times C$ distance matrix picks $2T$ distinct controls from the C controls available, with T controls picked by the first criterion, and T more controls picked by the second criterion. The two controls paired to rows 1 and $T + 1$ are the two controls for the first treated individual, . . ., and the two controls paired to rows i and $T + i$ are the two controls paired to treated individual i.

As discussed in [6], the method just described optimally splits and optimally pairs the C controls in the sense that the total of the $2T$ treatment-control distances is been minimized. In this specific sense, the method avoids duplication of controls in an optimal way.

As discussed in [6], optimal tapered matching may be used to form more than two comparison groups, or to select more than one individual from each comparison group. These tasks require only minor adjustments to the methods described here, for instance, stacking more than two distance matrices.

21.4 Building Matched Sets Using an Approximation Algorithm

21.4.1 A Near Optimal Solution to a Difficult Problem

In Sects. 21.2 and 21.3, a single reservoir of potential controls was divided to form two or more matched comparison groups with different attributes. In the current section, there are, at the outset, three well-defined groups and we wish to form matched triples, with one person from each group, while controlling for covariates.

As noted in Sect. 21.1, finding matched triples that minimize the total distance within triples is a computationally hard problem, but it is possible to construct approximation algorithms that run in polynomial time. Building upon the work of Yves Crama and Frits Spieksma [5], Bikram Karmakar et al. [7]. proposed an approximation algorithm for the computationally difficult problem of minimum distance matching with fine balance for three groups. The general algorithm [7] may be used with more than three groups, or with multiple controls drawn from each group, but it will be described here in the simplest case, namely finely balanced matched triples formed from three groups.

21.4.2 What Is Near-Fine Balance with Three Groups?

Given a nominal covariate with several or many categories, fine balance for matched pairs means that every category has the same frequency in the treated group and the matched control group; see Chap. 11. If there are K categories, there is a $2 \times K$ contingency table, treatment-versus-control by category, recording the number of treated or matched control individuals in each category. Fine balance means that the two counts in each column are equal. Fine balance constrains the marginal distribution of the K-category covariate, but it does not constrain who is paired to whom. A separate covariate distance is used to pair individuals who are close on the most important covariates.

With two groups, treatment or control, pair matching with fine balance is not always feasible. Pair matching is feasible whenever there are more controls in the control reservoir than there are treated individuals, but for fine balance to be feasible this pattern must hold in every category. Suppose that there are 10 treated individuals in the first category; however, the entire control reservoir contains only 9 controls in the first category. Then fine balance is not feasible in the first category. The total deviation from fine balance is the sum of the absolute differences in the counts in the first and second row of the $2 \times K$ contingency table. Fine balance means that total absolute deviation is zero. Near-fine balance means that the total absolute deviation is as small as possible. In the supposed situation, the total absolute deviation is at least 2, because the deficit of $1 = |10 - 9|$ in the first category must be offset by a surplus of 1 in some other category, resulting in $2 = 1 + 1$. A pair match is a minimum distance, near-fine match if it: (1) exhibits near-fine balance, (2) minimizes the total covariate distance among all pair matches that exhibit near-fine balance.

There is a second equivalent way to describe near-fine balance in matched pairs, and this second description generalizes to three or more groups. In addition to the $2 \times K$ contingency table describing the matched pairs—the match table—there is also a $2 \times K$ contingency table—the reservoir table—describing the treated group and the entire control reservoir. In the example in the previous paragraph, the second or reservoir table had, in the first column, a 10 in the first row and a 9 in the second row. In a near-fine match, the match table also had, in the first column, a 10 in the first row and a 9 in the second row, because that is as close as you can get to the desired (10, 10) counts for the first column. In the reservoir table, find the minimum of the two counts in each column, or 9 for the first column of the example. Near fine balance means that both counts in each column of the match table are greater than or equal to the minimum of the two counts in the corresponding column of the reservoir table. In the example, the minimum count in column one of the reservoir table was 9, and (10, 9) in the first column of the match table has both counts ≥ 9. We cannot have more than 9, but we want 10, so we settle for 9. In the match table for the matched pairs, the total count in the first row equals the total count in the second row. Therefore, if we minimize the absolute value of the deficits, $10 - 9 = -1 < 0$, then the surpluses take care of themselves.

The definitions of fine balance and near-fine balance now require only small adjustments for matched triples from three groups. There is now a $3 \times K$ match table and a $3 \times K$ reservoir table. If the three counts in each column of the match table are equal, then there is fine balance. If the three counts in each column of the match table are each greater than or equal to the minimum count in the corresponding column of the reservoir table, then there is near-fine balance. The algorithm of Karmakar et al. [7] achieves near-fine balance in this sense, so it achieves fine balance whenever fine balance is feasible.

21.4.3 An Approximation Algorithm

The idea behind the approximation algorithm will now be described. The algorithm requires covariate distances that satisfy the triangle inequality. This means that the distance between person i and person k is at most the sum of two other distances, namely the distance between person i and person j plus the distance between person j and person k. In words, i could travel to person k directly, or i could first visit j, then travel on to k, but going out of the way to visit j can only lengthen the trip. Many distances do satisfy the triangle inequality. Among distances that satisfy the triangle inequality are: the Euclidean norm, the Mahalanobis norm, and the Hamming distance. In matching, we often do not think about the triangle inequality when defining a covariate distance, but this is important in the approximation algorithm for matched triples. For detailed discussion of possible distances and penalties, see [7].

The three groups are ordered by the size of their reservoirs, smallest, middle, largest. Everyone in the smallest group is matched, so the number of matched triples equals the size of the smallest group. Near-fine balance is required for the three groups at once, as defined in the previous subsection.

There are two steps. First, a minimum distance, near-fine match is found between the smallest group and the middle group. That can be done in polynomial time using essentially the method in Chap. 11. The second step begins by computing a new distance matrix whose rows are the pairs, (i, j), from the first step and whose columns are the reservoir of controls, k, from the largest group. The distance between row (i, j) and column k is the sum of the two distances between i and k and between j and k. That is, k would be a good control for pair (i, j) if k is close to both i and j. The second step concludes by finding a minimum distance, near-fine match between the pairs (i, j), and the members k of the largest group. Again, that can be done in polynomial time using essentially the method in Chap. 11.

The two-step algorithm need not produce an optimal match, in the sense of minimizing the total distance within matched triples subject to the constraint of near-fine balance. We committed to the pair (i, j) before considering the consequences for k, so perhaps some of our (i, j) pairs are not the best beginnings for a matched triple. As noted in [5] and Sect. 21.1, finding the optimal match is not computationally feasible.

Although the two-step algorithm need not produce an optimal match, its solution cannot be very bad. Adapting ideas from [5], Karmakar et al. [7] show that the total distance from the two-step algorithm is at most twice the total distance of the unobtainable optimal solution. The proof of this [7, Proposition 1] uses the triangle inequality. Intuitively but very informally, if k is close to both i and j, then we may regret having paired i and j, but the triangle inequality limits the magnitude of our regret.

The method in [7] is more general than its description here. Only a limited aspect of the triangle inequality is required, and this has practical implications when defining the covariate distance. The method is not restricted to matched triples: it can match in a fixed ratio, say 1-to-2-to-4 for three groups yielding blocks of size $1 + 2 + 4 = 7$. The method can match more than three groups. The general method is implemented in Karmakar's `R` package `approxmatch`.

21.5 Is It Possible to Attenuate Unmeasured Bias?

21.5.1 Attenuation: Logically Possible But Small in Magnitude

Suppose that the observed covariate $\mathbf{x}$ is composed of two parts, $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2)$ where $\mathbf{x}_2$ is associated with treatment assignment Z but not with the potential outcomes (r_T, r_C) or unobserved covariates u.[1] Would it be wise to match for $\mathbf{x}_1$ but to leave $\mathbf{x}_2$ unmatched? This question has been debated [3, 20].

Those who argue in favor of leaving $\mathbf{x}_2$ unmatched argue that $\mathbf{x}_2$ creates no bias, yet introduces a random element in treatment assignment. You want to remove the bias from treatment assignment, not the random elements, or so this reasoning goes. Silly covariates aside, the most common argument against leaving $\mathbf{x}_2$ unmatched is that the required conditions make extensive reference to unobservable quantities, (r_T, r_C, u), so these conditions are speculative and difficult if not impossible to check. How can you be sure that $\mathbf{x}_2$ is unrelated to u?

Samuel Pimentel et al. [11] prove a qualitative result and supplement that with a quantitative evaluation. The qualitative result states that if the stringent conditions about $\mathbf{x}_2$ are true, then unmeasured bias from u is attenuated: the relevant value of Γ in Sect. 3.4 is strictly reduced [11, Proposition 1]. However, the quantitative evaluation says: Γ, though reduced, is not reduced by much. If the initial bias is enormous—$\Gamma = 10$, say—then there is meaningful bias attenuation—to perhaps $\Gamma = 9$—but the bias that remains is still enormous. If the initial bias is moderate in size—$\Gamma = 1.5$, say—then the attenuation is trivially small—$\Gamma = 1.47$. These numbers are quoted from [11, Table 1] where additional detail may be found. Pimentel et al. [11] conclude that too little bias attenuation is produced by assumptions that are too speculative.

[1] See [11, §2.1,§3.1] for the precise statement of these rather stringent conditions.

If leaving $\mathbf{x}_2$ unmatched takes a step in the right direction, but a step that is too small, then ask: Would forcing separation on $\mathbf{x}_2$ produce greater attenuation of unmeasured biases from u? Pimentel et al. [11, Table 2] show that forcing the treated group and the matched control group to differ substantially on $\mathbf{x}_2$ produces somewhat greater bias attenuation, but most of the unmeasured bias from u remains. Moreover, because the matched groups now differ dramatically in terms of $\mathbf{x}_2$, a small failure of the assumption that $\mathbf{x}_2$ is irrelevant could do substantial harm.

Is there a useful alternative to making speculative assumptions about $\mathbf{x}_2$ as the basis for leaving $\mathbf{x}_2$ unmatched?

21.5.2 Look More, Speculate Less

Should we speculate about the relationship between $\mathbf{x}_2$ and (r_T, r_C, u)? Should we decide to match for $\mathbf{x}_1$ instead of $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2)$ on the basis of speculation? Or should we put speculation aside, and look at the world from two perspectives? If we construct two comparison groups, one matched for $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2)$, the other matched for $\mathbf{x}_1$, perhaps forcing separation on $\mathbf{x}_2$, then we will see what there is to be seen. Theory teaches us where to look, how to look; it is not a substitute for looking.

A simple approach [11] is to construct matched triples, perhaps overlapping triples as in Sect. 21.2, or perhaps nonoverlapping triples as in Sect. 21.3, where one control in each triple is matched to the treated individual for $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2)$, and the other control is matched for $\mathbf{x}_1$. For the second control, there is the option of leaving $\mathbf{x}_2$ unmatched or forcing separation on $\mathbf{x}_2$. In this way, the investigator sees both analyses, can and must interpret and reconcile both analyses; however, the investigator need not endorse or depend upon speculative assumptions about $\mathbf{x}_2$.

The second controls, dissimilar on $\mathbf{x}_2$, are called prodded controls [11, §3.1]. The thought is that $\mathbf{x}_2$ has prodded some of these controls to abstain from receiving the treatment. Often, the prodded controls are available for free; often, they are far from treated individuals on $\mathbf{x}_2$, so a sample matched for $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2)$ would have excluded them anyway. Looking at the prodded controls often entails looking at available data that otherwise might be ignored.

Using two controls in this way requires a plan to control the error rate when two or more analyses are performed; see [11, §6] and Chap. 23 for several possible plans. If one control group is given priority when "testing in order," then there is no correction for multiple testing when the high-priority group is considered, so again the low-priority control group is available for free. If the control groups are combined in the first step of "testing in order," then the power may be increased by the increased sample size, with the two control groups considered separately in the second step. For specifics, see [11, §6], [14] and Sect. 23.3.

21.5.3 Smoking and Elevated Homocysteine Levels

An interesting study by Bazzano et al. [2] used data from the National Health and Nutrition Examination Survey (NHANES) to ask whether cigarette smoking causes an increase in homocysteine levels in the blood. Elevated homocysteine levels are thought to be a risk factor for various diseases, including cardiovascular disease. As an illustration, Pimentel et al. [11] conducted a parallel analysis using a more recent NHANES survey that measured homocysteine, namely NHANES 2005–2006. This parallel analysis illustrates the joint use of matched and prodded controls.

Bazzano et al. [2] had adjusted for a number of directly biological covariates possibly associated with homocysteine levels, including age, gender, and body-mass index (BMI); call these $\mathbf{x}_1$. They did not emphasize adjustments for socioeconomic covariates, such as education and income; call these $\mathbf{x}_2$. In the USA today, cigarette smoking is relatively uncommon among individual with relatively more education and income. Would leaving $\mathbf{x}_2$ unmatched attenuate bias or introduce bias? Should we compare biologically similar smokers and controls, even if the controls have more education and income? Should we force separation on $\mathbf{x}_2$ in the hope of producing greater attenuation of unmeasured bias? Or would forced separation only enlarge the bias from $\mathbf{x}_2$?

Why do some people smoke while others do not? Among people in the USA with less income and less education, smoking appears to be an individual choice: many such people smoke, even more do not. Among people with more income and more education, the decision not to smoke is far more uniform: relatively few such people smoke. Perhaps people with more education make wiser choices, or perhaps the social stigma associated with smoking overwhelms them.

To affect homocysteine levels, education and income would have to act through some mechanism with biological consequences. Such mechanisms may exist. Education and income may have consequences for diet, exercise, and the quality of health care. Is it safe to leave $\mathbf{x}_2$ unmatched? Should we speculate and argue? Or should we look and see? Rather than speculate, why not look at the situation from both perspectives. Because there are relatively few smokers with high income and education, the control group that does not match for $\mathbf{x}_2$ is mostly built from potential controls who would have been otherwise discarded.

Smokers smoked every day for the last 30 days, and smoked an average of at least 10 cigarettes on each day. Nonsmokers did not smoke in the last 30 days and smoked fewer than 100 cigarettes in their lives.

Table 21.1 describes numeric covariates, displaying their eighths, quartiles, and medians. The numeric covariates are age, body-mass index (BMI), education, and income. Education has five numeric categories, 1 for less than 9th grade, 3 for high school, 5 for BA degree or higher. Income is the ratio of family income to the poverty level, capped at 5 by NHANES. In addition, there were three binary covariates: gender, black race, and Hispanic. Here, $\mathbf{x}_1$ contains age, BMI, gender, black race, and Hispanic, while $\mathbf{x}_2$ contains education and income. Finally, Table 21.1 displays the outcome, the log of homocysteine in blood plasma recorded as μmol/l.

Table 21.1 Does smoking increase homocysteine?

	Quantile				
	1/8	1/4	1/2	3/4	7/8
	Covariates				
	Age in $\mathbf{x}_1$				
Smoker	26	34	45	57	64
Matched	25	32	45	59	70
Prodded	28	33	44	57	65
	BMI in $\mathbf{x}_1$				
Smoker	20.7	22.7	26.2	31.1	35.7
Matched	21.7	23.2	26.4	31.1	36.2
Prodded	22.0	24.2	26.8	31.5	35.4
	Education in $\mathbf{x}_2$				
Smoker	2	2	3	4	4
Matched	2	2	3	4	4
Prodded	4	4	5	5	5
	Income)/poverty in $\mathbf{x}_2$				
Smoker	0.7	1.1	2.0	3.6	4.9
Matched	0.8	1.1	2.1	3.8	5.0
Prodded	2.3	3.5	5.0	5.0	5.0
	Outcome				
	log(homocysteine)				
Smoker	6.1	6.8	8.4	10.4	11.9
Matched	5.6	6.4	7.5	9.3	10.9
Prodded	5.3	6.2	7.6	9.2	10.7

Smokers and controls in 512 matched triples. "Matched" controls were matched for both biological covariates $\mathbf{x}_1$ and socioeconomic covariates $\mathbf{x}_2$. "Prodded" control was matched for biological covariates $\mathbf{x}_1$, but pushed to have higher levels of socioeconomic covariates $\mathbf{x}_2$, namely education and income. Note carefully that prodded controls have higher education and income than treated subjects

In Table 21.1, the matched controls are similar to smokers for all of $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2)$. The prodded controls are similar to smokers in terms of $\mathbf{x}_1$, but they are forced to have much higher education and income, $\mathbf{x}_2$. Forced separation can be introduced by using fine balance backwards, constraining the match to be imbalanced rather than balanced [11]. Alternatively, directional penalties added to distances can encourage imbalances [29].

Notably, in Table 21.1, the logs of homocysteine levels are higher for smokers than for both matched and prodded controls. The higher income and education of prodded controls seems to matter little, if at all. The smoker-control comparison is insensitive to a bias of $\Gamma = 1.5$, with an upper bound on the P-value ≤ 0.05, whether matched or prodded controls are used. Here, $\Gamma = 1.5$ is a bias that

doubles the odds of smoking and quadruples the odds of elevated homocysteine; see Sect. 3.6. See Pimentel et al. [11, Table 4] for specifics and for a planned sensitivity analysis of both groups controlling the family-wise error rate.

Campbell [4] wrote: "Clarity of inference sometimes may be improved by deliberately reducing the quality of part of the data." Although one might have more confidence in the controls matched for $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2)$ in Table 21.1, there is nonetheless something reassuring in seeing that the answer is unchanged when using better educated, wealthier controls.

21.6 Further Reading

The exterior match is discussed in [19] and applied in [21]. Optimal tapered matching was proposed by Shoshana Daniel et al. [6]. Approximation algorithms for matching several groups are discussed by Bikram Karmakar et al. [7]; see also [5]. For general discussion of approximation algorithms, see [25, 27]. The method and homocysteine example in Sect. 21.5 are from Samuel Pimentel et al. [11].

21.7 Software

Given matched pairs from two entwined matched pair control groups, the `R` package `exteriorMatch` removes the duplicates, minimally adjusts the pairing, and returns the exterior match discussed in Sect. 21.2.

Bikram Karmakar's `R` package `approxmatch` can produce a match with near-fine balance for three or more groups using the approach described in Sect. 21.4.

Bibliography

1. Baiocchi, M., Small, D.S., Lorch, S., Rosenbaum, P.R.: Building a stronger instrument in an observational study of perinatal care for premature infants. J. Am. Stat. Assoc. **105**, 1285–1296 (2010)
2. Bazzano, L.A., He, J., Muntner, P., Vupputuri, S., Whelton, P.K.: Relationship between cigarette smoking and novel risk factors for cardiovascular disease in the United States. Ann. Int. Med. **138**, 891–897 (2003)
3. Brooks, J.M., Ohsfeldt, R.I.: Squeezing the balloon: propensity scores and unmeasured covariate balance. Health Serv. Res. **48**, 3078–3094 (2013)
4. Campbell, D.T.: Prospective: artifact and control. In: Rosenthal, R., Rosnow, R. (eds.) Artifact in Behavioral Research, pp. 351–382. Academic, New York (1969)
5. Crama, Y., Spieksma, F.C.: Approximation algorithms for three-dimensional assignment problems with triangle inequalities. Eur. J. Oper. Res. **60**, 273–279 (1992)

6. Daniel, S., Armstrong, K., Silber, J.H., Rosenbaum, P.R.: An algorithm for optimal tapered matching, with application to disparities in survival. J. Comput. Graph. Stat. **17**, 914–924 (2008)
7. Karmakar, B., Small, D.S., Rosenbaum, P.R.: Using approximation algorithms to build evidence factors and related designs for observational studies. J. Comput. Graph. Stat. **28**, 698–709 (2019)
8. Kelz, R.R, Sellers, M.M., Niknam, B.A., Sharpe, J.E., Rosenbaum, P.R, Hill, A.S., Zhou, H., Hochman, L.L., Bilimoria, K.Y., Itani, K., Romano, P.S., Silber, J.H.: A national comparison of operative outcomes of new and experienced surgeons. Ann. Surg. (2019, to appear). https://doi.org/10.1097/SLA.0000000000003388
9. Lu, B., Rosenbaum, P.R.: Optimal matching with two control groups. J. Comput. Graph. Stat. **13**, 422–434 (2004)
10. Meyer, B.D.: Natural and quasi-experiments in economics. J. Bus. Econ. Stat. **13**, 151–161 (1995)
11. Pimentel, S.D., Small, D.S., Rosenbaum, P.R.: Constructed second control groups and attenuation of unmeasured biases. J. Am. Stat. Assoc. **111**, 1157–1167 (2016)
12. R Development Core Team.: R: A Language and Environment for Statistical Computing. R Foundation, Vienna (2007). http://www.R-project.org
13. Rosenbaum, P.R.: From association to causation in observational studies. J. Am. Stat. Assoc. **79**, 41–48 (1984)
14. Rosenbaum, P.R.: The role of a second control group in an observational study (with Discussion). Stat. Sci. **2**, 292–316 (1987)
15. Rosenbaum, P.R.: On permutation tests for hidden biases in observational studies. Ann. Stat. **17**, 643–653 (1989)
16. Rosenbaum, P.R.: Differential effects and generic biases in observational studies. Biometrika **93**, 573–586 (2006)
17. Rosenbaum, P.R.: Testing hypotheses in order. Biometrika **95**, 248–252 (2008)
18. Rosenbaum, P.R.: Using differential comparisons in observational studies. Chance **26**(3), 18–23 (2013)
19. Rosenbaum, P.R., Silber, J.H.: Using the exterior match to compare two entwined matched comparison groups. Am. Stat. **67**, 67–75 (2013)
20. Sani, A., Groenwold, R.H.H., Klungel, O.H.: Propensity score methods and unobserved covariate balance. Health Serv. Res. **49**, 1074–1082 (2014)
21. Silber, J.H., Rosenbaum, P.R., Clark, A.S., Giantonio, B.J., Ross, R.N., Teng, Y., Wang, M., Niknam, B.A., Ludwig, J.M., Wang, W., Even-Shoshan, O., Fox, K.N.: Characteristics associated with differences in survival among black and white women with breast cancer. J. Am. Med. Assoc. **310**, 389–397 (2013)
22. Silber, J.H., Rosenbaum, P.R., Ross, R.N., Niknam, B.A., Ludwig, J.M., Wang, W., Clark, A.S., Fox, K.N., Wang, M., Even-Shoshan, O., Giantonio, B.J.: Racial disparities in colon cancer: a matched cohort study. Ann. Int. Med. **161**, 845–854 (2014)
23. Silber, J.H., Rosenbaum, P.R., Ross, R.N., Niknam, B.A., Hill, A.S., Bongiorno, D., Even-Shoshan, O., Fox, K.N.: Disparities in breast cancer survival by socioeconomic status despite medicare and medicaid insurance. Milbank Q. **96**, 706–754 (2018)
24. Stuart, E.A., Rubin, D.B.: Matching with multiple control groups with adjustment for group differences. J. Educ. Behav. Stat. **33**, 279–306 (2008)
25. Vazirani, V.V.: Approximation Algorithms. Springer, New York (2010)
26. Walker, A.M.: Matching on provider is risky. J. Clin. Epidemiol. **66**, 565–568 (2013)
27. Williamson, D.P., Shmoys, D.B.: Design of Approximation Algorithms. Cambridge University Press, New York (2011)
28. Yang, F., Zubizarreta, J.R., Small, D.S., Lorch, S., Rosenbaum, P.R.: Dissonant conclusions when testing the validity of an instrumental variable. Am. Stat. **68**, 253–263 (2014)
29. Yu, R., Rosenbaum, P.R.: Directional penalties for optimal matching in observational studies. Biometrics (2019, to appear). https://doi.org/10.1111/biom.13098

Part V
Planning Analysis

Chapter 22
After Matching, Before Analysis

Abstract Three design tasks may usefully follow matching and precede planning of the analysis. Splitting the sample of I pairs into a small planning sample and a large analysis sample may aid in planning the analysis in a manner that increases the design sensitivity. If there will be analytic adjustments for some unmatched variables, it is prudent to check that the matched samples exhibit sufficient overlap on unmatched variables to permit analytic adjustments. Quantitative analysis of matched samples may usefully be combined with qualitative examination and narrative description of a few closely matched pairs.

22.1 Split Samples and Design Sensitivity

In Chaps. 15–18, the sensitivity of the design of an observational study to unmeasured biases, $\widetilde{\Gamma}$, was found to depend upon many things: unit heterogeneity, coherence among outcomes, doses, uncommon but dramatic responses to treatment, the strength of instrumental variables, among others. In some contexts, these considerations may be difficult to evaluate during design, prior to the examination of outcomes. An interesting option with good properties entails splitting the sample at random into a small planning sample and a large analysis sample, making some design decisions based on the planning sample, applying those decisions in the analysis sample. This section discusses the matter in brief outline, based upon a paper by Ruth Heller et al. [17].

Sample splitting is most familiar from cross-validation [42]. In cross-validation, a sample is split, usually at random, and a prediction is developed in one portion and tested in the other. Often the process is repeated with many random splits in an effort to develop an honest appraisal of the performance of some method of prediction.

Less familiar is the use of sample splitting to decide what question to ask, which hypothesis to test. David Cox [8] examined the use of split samples as an alternative to corrections for multiple testing in randomized experiments, such as corrections based on the Bonferroni inequality. Instead of testing many hypotheses, the sample is split, the first portion used to select a promising hypothesis, the second portion used to test that one hypothesis. In randomized experiments, Cox found that splitting

© Springer Nature Switzerland AG 2020

P. R. Rosenbaum, *Design of Observational Studies*, Springer Series in Statistics,
https://doi.org/10.1007/978-3-030-46405-9_22

was slightly inferior in terms of power to corrections for multiple testing in simple stylized settings. Cox noted the flexibility of splitting, which permits confirmation of unanticipated exploratory findings.

Sample splitting has much better properties in observational studies than in randomized experiments, outperforming, for example, the Bonferroni inequality [17]. In an idealized randomized experiment, there is no bias from unmeasured covariates, and the power of a conventional test is properly the focus of concern. In contrast, in an observational study, bias from failure to adjust for an unmeasured covariate is an ever-present worry, and the power of a sensitivity analysis is the relevant concern. As seen in Chap. 15, in favorable situations in which there is an effect but no unmeasured bias, as the sample size I increases, $I \to \infty$, the power of a sensitivity analysis tends to 1 for all $\Gamma < \widetilde{\Gamma}$ and to 0 for all $\Gamma > \widetilde{\Gamma}$, where $\widetilde{\Gamma}$ is the design sensitivity [27]. Again, the design sensitivity $\widetilde{\Gamma}$ depends upon many aspects of the design; see Chaps. 15–18 and [27–30, 41]. Suppose that one could exchange a portion of the sample, say $(1-\varrho)\ I$ observations with $0 < \varrho < 1$, for a larger design sensitivity, say $\widetilde{\Gamma}^* > \widetilde{\Gamma}$, in the remaining portion. That is, suppose that a "planning sample" of, say, $10\% = (1-\varrho)$ of the observations led to better decisions about design, which in turn raised the design sensitivity, from $\widetilde{\Gamma}$ to $\widetilde{\Gamma}^*$, in the remaining 90%, the "analysis sample." If the sample size, I, is large, this is a wonderful swap. After all, as $I \to \infty$, without the split, the power is tending to zero for Γ in the open interval $\left(\widetilde{\Gamma},\ \widetilde{\Gamma}^*\right)$, whereas with the split the power is tending to one on the same interval. Admittedly, this is a limiting argument about a large sample, but as intuition might suggest and detailed calculations in [17] confirm, it does not take a very large sample size I for a larger design sensitivity, $\widetilde{\Gamma}^*$ instead of $\widetilde{\Gamma}$, to overwhelm a small reduction in sample size, to $\varrho\ I$ from I.

The discussion in the previous paragraph presumed that a planning sample of size $(1-\varrho)\ I$ could raise the design sensitivity in the analysis sample of size $\varrho\ I$. This presumption has two aspects, both more important than the loss of $(1-\varrho)\ I$ observations. The first aspect is that there must be design decisions that would increase $\widetilde{\Gamma}$ if those decisions were correctly made. Concerning the first aspect, Chaps. 15–18 offer some hope but no guarantees. The second aspect is that the planning sample must be large enough to lead to correct decisions about design when there are correct decisions to be made. Numerical results relevant to the second aspect are given in [17]. In several cases, $(1-\varrho)\ I = 100$ planning observations with $\varrho\ I = 900$ analysis observations worked extremely well; see [17] for specifics and detailed recommendations. Arguably, one needs a planning sample of size $(1-\varrho)\ I$ adequate to make correct design decisions if there are substantial gains in design sensitivity to be had, with little to be gained by increasing the size of the planning sample beyond that.

The practical example in [17] seemed too small to hope for much. There were only $I = 36$ pairs, with $(1-\varrho)\ I = 6$ used for planning and $\varrho I = 30$ used for analysis. As an illustration, that analysis was repeated with 30 independent splits. In every one of the 30 splits, the six planning observations correctly guided the one important design decision that materially affected the sensitivity to unmeasured

biases. That is too much to expect in general from six observations, but in larger studies, sample splitting has much to offer.

A much larger practical example of sample splitting in the design of an observational study is discussed by Zhang et al. [46]. An alternative approach to sample splitting is discussed by Zhao et al. [47].

22.2 Are Analytic Adjustments Feasible?

In an observational study, matching sets up certain comparisons, sets up a certain analogous experiment: it builds adjustments for observed covariates into the design of an observational study. In Mervyn Susser's phrase, matching "simplifies the conditions of observation," aiding transparency (see Chap. 6).

Adjustments built into the design cannot be easily removed later. Under certain circumstances, one may plan two analyses, one with an adjustment for a certain variable, the other without that adjustment. In this case, the variable is not controlled by matching but is controlled in the statistical analysis. An aspect of the design of an observational study entails verifying that both analyses are feasible. This means checking that an unmatched variable exhibits sufficient overlap in treated and control groups that analytical adjustments are an interpolation, not an extrapolation or a fabrication.

Typically, one leaves unmatched a variable of ambiguous status. In that way, different analyses may assign the variable different roles [6]. If a concomitant variable is not truly a covariate, if it may have been affected by the treatment, then an adjustment for it may introduce a bias that would not otherwise have been present [22]. That said, a variable may have an ambiguous status, perhaps slightly affected by the treatment, but plausibly standing in as a surrogate for an important covariate that was not measured. For instance, two studies in 1982, by James Coleman et al. [7] and Arthur Goldberger and Glen Cain [16], asked whether Catholic and other private high schools did more than public schools to increase the cognitive test performance of their students. Important covariates in this context are cognitive test scores before high school, but these were not available. Both studies make various comparisons of sophomore and senior test scores, with the view that sophomore test scores, though possibly affected by public-vs.-private schooling, might serve in one way or another as a proxy for the unmeasured covariates of test scores prior to high school. An important variable with an ambiguous status does not necessarily imply an ambiguous conclusion: perhaps the conclusion is unchanged by changing the status of the variable. If a variable of ambiguous status is left unmatched, analyses may examine issues of this sort.

Another type of variable of ambiguous status exhibits "seemingly innocuous confounding." That is, the covariate predicts a substantial fraction of the variation in treatment assignment, but seems fairly innocuous, without obvious importance to the outcomes under study. At the same time that the treatment is observed to be confounded with something innocuous, it may also be confounded

with unmeasured covariates that are not innocuous. In certain circumstances, the innocuous confounding—the devil you know, in the words of the proverb—may seem less worrisome than confounding with unmeasured covariates—the devil you don't.[1] Should "seemingly innocuous confounding" be removed? Perhaps. Perhaps it is not innocuous, and only seems so. And yet, if all the innocuous confounding is removed, then perhaps the haphazard aspect of treatment assignment is removed as well, and the remaining aspects of treatment assignment are governed by biases from relevant unmeasured covariates. Again, when a variable has an ambiguous status, conclusions may be firmer if shown to be untouched by this ambiguity. Consider an example.

Recall from Sect. 13.3 the study by Jeffrey Silber and colleagues [33, 40] of slightly earlier or slightly later discharges of 1402 premature infants from five neonatal intensive care units (NICUs) in the Northern California Kaiser-Permanente Medical Care Program between 1998 and 2002. The babies were paired into 701 pairs so that the two babies looked similar in terms of measured (time-dependent) covariates on the day the first baby, the early baby, was discharged, while the other baby, the late baby, stayed in the NICU for a few more days, growing older and more mature. Some of the variation in discharge day seems to have been governed by the day of the week: reach maturity on Thursday and you go home on Friday; reach maturity on Saturday and you go home on Tuesday, or so it appeared. There were five hospitals in the Northern California Kaiser-Permanente Medical Care Program that provided data for this study, and hence five NICUs, one at each hospital. Some of the variation in discharge day seems to reflect styles of practice at the different NICUs: some are faster than others. You could make a case, of course, that these sources of variation are not innocuous. You could make a case that the baby discharged Friday morning was judged healthier by the attending neonatologist than the baby held until Tuesday, observed covariates to the contrary not withstanding. Although these were all Kaiser-Permanente moms in hospitals run by Kaiser-Permanente, you could make a case that the moms at certain hospitals or the hospitals themselves were not really the same. Counter to this, you could make a case that the variation in discharge day predicted by the identity of the NICU or the day of the week is less worrisome than unexplained variation. Should we speculate and argue, or should we look and see? Let's look and see.

Table 22.1 shows the distribution of the 701 pairs of two babies at the five NICUs. The table counts pairs, not babies, so the total count is 701 pairs, not 1402 babies. Of the 701 pairs, 233 fall on the diagonal: the early and late baby are in the same NICU. Some NICUs discharge earlier than others: NICU C contributed more late babies than early babies, while NICU D did the opposite. On the other hand, despite the trend, there are many early babies from NICU C, specifically 95, and many late babies from NICU D, specifically 89. The variation between NICUs in discharge rate accounts for part but not all of the variation in discharge day; therefore, the

[1] Erasmus: "Nota res mala optima (an evil thing known is best)" [10, #85, p. 123].

Table 22.1 Pairs classified by the neonatal intensive care unit (NICU, A–E) of the early baby and the late baby for 701 matched pairs

Early baby	Late baby A	B	C	D	E	Total	Percent	Late/early ratio
A	18	23	31	13	29	114	16	0.79
B	20	65	24	17	36	162	23	0.93
C	10	19	30	12	24	95	14	1.69
D	26	24	41	36	38	165	24	0.54
E	16	19	35	11	84	165	24	1.28
Total	90	150	161	89	211	701		
Percent %	13	21	23	13	30		100	

The table counts pairs, so the total count is 701. NICU C had 161 late babies and 95 early babies, for a Late/Early ratio of $161/95 = 1.69$, while NICU D had 89 late babies and 165 early babies, for a Late/Early ratio of $89/165 = 0.54$. Percentages are rounded

variation due to the NICU can be removed analytically. In this way, we will see the situation both with innocuous confounding and without.

Table 22.1 is part of the design. There is no information about outcomes in Table 22.1. The pattern in Table 22.1 reveals that it is safe to leave NICU unmatched; within these 701 pairs matched for many characteristics of baby and mom, there is enough overlap in the NICUs to permit adjustments. We may perform two analyses, one that ignores the NICU, viewing the NICU as an innocent source of variation in the day of discharge, the other analysis removing entirely the variation in discharge day that can be predicted from the NICU. Deciding whether to match for the NICU is part of the design, and Table 22.1 is part of that decision.

As will be discussed in detail in Sect. 23.6, one of the outcomes in Silber et al.'s [33, 40] study concerned sick-baby services in an interval approximately 6 months long following the discharge of the late baby, when both babies were home. Did early discharge prompt a greater need for sick-baby services after discharge? Well-baby services, such as checkups, were not included. Sick-baby costs include both minor events with minor costs, and major events, such as readmission to the NICU, with enormous costs. Services were converted to dollar amounts according to a fixed schedule, so that babies who received the same services had the same dollar score for those services. Here, postdischarge sick-baby cost is serving as a measure of the severity of the health problem, for which it is not a bad proxy. See Sect. 23.6 and [33, 40] for specifics and additional outcomes.

Table 22.2 is a letter value display [44] of the 701 matched pair differences, late-minus-early, in postdischarge sick-baby costs. Five deaths are recorded as infinite costs, three among the early babies, two among the late babies; see Sect. 23.6 and [40] for more about the deaths. The median difference in cost is \$5 and the average of the quartiles is \$32. The distribution is extremely long-tailed relative to the Normal distribution, as is seen from the pseudo-spreads in Table 22.2, which would be approximately constant for a very large sample from a Normal distribution. There is no sign of increased postdischarge costs for early babies.

Table 22.2 Letter value display of difference in postdischarge sick-baby costs in dollars in 701 pairs of a late and an early baby

Quantile %	'Letter'	Lower	Upper	Mid	Spread	Pseudo-spread
50	1/2	5	5	5	0	
25	1/4	−223	287	32	510	378
12.5	1/8	−1063	2313	625	3376	1467
6.25	1/16	−4089	4956	433	9045	2948
3.125	1/32	−9893	12,944	1526	22,836	6130

The difference is late-minus-early. Various quantiles are displayed, together with their averages (the mids), differences (the spreads), and the ratio of the spreads to the corresponding spreads of the standard Normal distribution. Five deaths are recorded as infinite costs, three among the early babies, two among the late babies; see Sect. 23.6 for detailed discussion of the deaths

One analysis ignores the NICU. It allows the variation in practice style at different NICUs to influence the day of discharge. A baby who happened to be born in NICU C might stay a few extra days, and one born in NICU D may be discharged a few days early. For two babies with the same matched covariates in Sect. 13.3, this analysis views the NICU as irrelevant to outcomes except to the extent that it accelerates or retards discharge. Ignoring the NICU, procedures associated with Wilcoxon's signed rank statistic (Sect. 2.4) give a point estimate of an additive effect τ of \$17 and a 95% confidence interval of $\left[-\$20,\ \$56\right]$; see Sect. 23.6 for additional analyses ignoring the NICU, including a sensitivity analysis.

The analysis using Wilcoxon's statistic viewed all 2^{701} possible treatment assignments in the 701 pairs as equally probable. Table 22.1 suggests they are not equally probable; rather, they are predictable using the NICU. For instance, look at row D and column C of Table 22.1 in comparison with row C and column D. In those two cells of Table 22.1 are $41 + 12 = 53$ pairs of two babies, one born at NICU C, the other at NICU D. In those 53 pairs, the odds that the early baby is at NICU D are estimated at $41/12 = 3.4$ to 1. The alternative analysis ([24] and [26, §3]) does not consider all 2^{701} possible treatment assignments, but rather a subset of treatment assignments that exhibits the same imbalance as observed in Table 22.1. For instance, in the 53 pairs with one baby at NICU C and another at NICU D, Wilcoxon's statistic considers all $2^{53} = 9 \times 10^{15}$ possible treatment assignments as equally probable and uses them all, but the alternative procedure considers only the $53!/\,(41!\,12!) = 2.7 \times 10^{11}$ treatment assignments that keep 12 early babies in NICU C and 41 late babies in NICU C. For the 233 pairs on the diagonal in Table 22.1, all possible treatment assignments are considered, but for the off-diagonal pairs, the observed imbalance is preserved. The distribution of Wilcoxon's statistic, or any

other statistic, is compared to this restricted or conditional permutation distribution.[2]

The alternative analysis that adjusts for the NICU yields a point estimate of an additive effect τ of $-\$6$ and a 95% confidence interval of $[-\$36,\ \$53]$. Some of the variation in day of discharge can be predicted from the NICU, a seemingly innocuous form of confounding. Two analyses were performed, one that ignored the predictable variation, the other that removed the predictable variation. The two analyses produce similar inferences about τ. Although the status of the NICU is ambiguous, the conclusion is less so. Indeed the pair of analyses provides some reassurance: different sources of variation in the day of discharge produce similar estimates of the effect of slightly delayed discharge.

Is it safe to leave a covariate unmatched? Will analytic adjustments be feasible for that covariate? Is there sufficient overlap on an unmatched covariate? Will analytic adjustments require an interpolation, not an extrapolation or a fabrication? In the design of an observational study, a table such as Table 22.1 recording treatment assignment and unmatched covariates provides answers to these questions.

22.2.1 Tapered Matching and the Exterior Match

An alternative to leaving an ambiguous covariate unmatched is to construct matched triples, one treated subject and two controls, so that one control is matched for the covariate and the other is not. This ensures that both analyses are feasible, and provides a formal test as to whether their answers differ, as well as a confidence statement about the magnitude of the difference. This topic was discussed in Sects. 21.2, 21.3 and 21.5 and in [9, 19, 21, 34].

22.3 Matching and Thick Description

22.3.1 Thick Description

Matching compares groups of treated and control subjects, often pairs of treated and control subjects, who look comparable in terms of measured covariates $\mathbf{x}_{ij}$. A constant worry is that subjects who look comparable may not be comparable,

[2]When is this analysis appropriate? Suppose the NICUs are included among the observed covariates and after including the NICUs, (3.5)–(3.8) is then true. Suppose further that the NICUs enter the propensity score as additive constants on the logit scale. Then the conditioning just described eliminates the additive NICU parameters, leaving behind a known distribution of treatment assignments that gives the remaining assignments equal probabilities. The distribution of the Wilcoxon statistic is then approximated using its easily computed moments. See [24] and [26, §3] for the paired case, as here, and see [23, 25] and [26, §3] for the general case.

that they look comparable only because important distinctions are omitted in the observed covariates $\mathbf{x}_{ij}$, distinctions that are to be found only in unobserved covariates u_{ij} for which subjects are not matched.

In matching, people remain intact as people. In contrast, in model-based adjustments, people recede into the background to be replaced by features of a model. Where there were once people, there are now parameters in a model. A relatively neglected but potentially important aspect of matching is the possibility of taking a close look at a few well-matched pairs of people, providing a narrative account of these pairs.[3]That is, one might provide a "thick description" of a small subset of pairs. Jeffrey Silber and I tried this in a study of mortality after surgery in the Medicare population in Pennsylvania, and the current section is a brief summary of the resulting paper [32].

It is a staple of professional education—in medicine, in business, in law—that a narrative account of a few well-chosen cases is often enlightening and instructive. For instance, in this spirit, the *New England Journal of Medicine* often publishes its "Case Reports from the Massachusetts General Hospital"; see also the essay by Jan Vandenbroucke [45]. The professions do not believe that broad generalizations can be warranted by examination of a few cases; rather, they believe that to become acquainted with generalizations without engaging particular cases to which those generalizations are supposed to apply is to run the risk of developing a fairly scholastic perspective.

In certain social sciences, there is a division of labor, with some investigators, sometimes called "qualitative researchers," creating narrative accounts from a few cases, and other investigators, sometimes called "quantitative researchers," creating analyses from data describing many people. Matching provides a framework within which qualitative and quantitative research can usefully interact within a single investigation. A narrative account of a few pairs could go hand in hand with a statistical analysis of many pairs.

22.3.2 What Is Thick Description?

The term "thick description" was introduced by a philosopher, Gilbert Ryle [35, p. 479], and was made popular by an anthropologist, Clifford Geertz [13]; see [32,

[3]As has been discussed several times, modern matching methods try to balance high-dimensional observed covariates in treated and control groups, viewing close individual matched pairs as a secondary concern. In light of this, treated and control groups may be balanced with respect to observed covariates using many pairs that are not individually close on observed covariates. The methodology discussed in the current section is most appropriate for the subset of matched pairs in which the individual pairs are closely matched. Typically, there will be many such pairs, even if many other pairs contribute to covariate balance without being closely matched as pairs with respect to covariates. If a distance, such as the Mahalanobis distance, is used in matching (see Chap. 9), then a closely matched pair will be recognizable by a small distance within the pair.

pp. 221–222] for quotes from Ryle and Geertz that give a brief sense of the nature of their concerns. The aim of thick description according to Geertz [14, p. 152] is "to render obscure matters intelligibly by providing them with an informing context" [14, p. 152]. The importance of an informing context is conveyed by the following passage by Howard Becker [2, p. 58]:

> [I]f we don't find out from people what meanings they are actually giving to things, we will still talk about those meanings. In that case, we will, of necessity, invent them, reasoning that the people we are writing about must have meant this or that, or they would not have done the things they did. But it is inevitably epistemologically dangerous to guess at what could be observed directly. The danger is that we will guess wrong, that what looks reasonable to us will not be what looked reasonable to them. This happens all the time, largely because we are not those people and do not live in their circumstances.

It is quite easy to misinterpret data from a large electronic data set [3]. The attempt to create a narrative account of a few cases may serve to test interpretations of data and the concepts in which those interpretations are cast. An example follows.

22.3.3 Example: Mortality After Surgery

Silber et al. [37–39]'s study of mortality after surgery used (inexpensive) administrative data to create a matched sample from the Medicare population in Pennsylvania, which was followed by (expensive) chart abstraction for the matched sample. In this study, deaths shortly following surgery were matched to ostensibly similar survivors based on administrative data from Medicare. Although the best data would come from chart abstraction for the matched sample, it was important to have a reasonable match based on administrative data, so that charts for suitable patients would be abstracted.

In [32], a preliminary matched sample was constructed, and Medicare records for a handful of well-matched pairs were contrasted with far more detailed information in the hospital chart. The preliminary match looked reasonable in terms of the measured variables from Medicare's administrative records. However, even though only a few cases were examined, it was immediately clear from chart data that the initial match was not adequate and that many definitions that we created to interpret the administrative records required revision. Many of the mistakes we unearthed were obvious mistakes when viewed from the step-by-step narrative perspective of a hospital chart, but were not obvious from the perspective of computerized administrative records, or at least not obvious to us.

For instance, we generally wanted to avoid matching on events that were subsequent to surgery, because these might be outcomes of surgery or of subsequent care. Although that principle is sound, our initial use of it was mechanical and thoughtless. For instance, if surgery reveals that a patient has metastatic cancer, that cancer existed prior to the surgery that discovered it. Metastatic cancer is, of course, important to the patient's prognosis and to proper interpretation of the care the patient subsequently received. In light of this, we changed our definitions related

to cancer and several other medical conditions that could not have begun during a hospital stay. No one would make this mistake faced with a narrative account describing an individual patient, but it is less obvious in computerized administrative records, and it might not be easily discerned from the coefficients of a model. If some mistakes are difficult to make when a situation is viewed from a narrative perspective, then that constitutes a good reason for expending the time and effort required to view the situation from a narrative perspective.

We also redefined "died shortly after surgery." One of the examined charts described a patient who had "survived after surgery" by our original definition, but after a long stay had died without leaving the hospital. This case prompted us to reexamine the administrative data, only to discover that the fate of this one patient was sufficiently common to merit a different definition of "died shortly after surgery."

For additional examples and further discussion, see [32]. The example in [32] is a fairly limited use of matching in conjunction with narrative description, but extensive use is possible. For three compelling examples of extended narrative accounts (without matching), see [1, 4, 12].

22.4 Further Reading

Section 22.1 is a brief outline of a paper by Ruth Heller et al. [17]. A large observational study designed with the aid of a split sample is discussed by Kai Zhang et al. [46]. An alternative approach to sample splitting, the so-called cross-screening, is discussed by Qingyuan Zhao et al. [47]. Ashkan Ertefaie et al. [11] use split samples when strengthening an instrument, as discussed in Sect. 5.4.

The method of adjustment in Sect. 22.2 is from [24]. Although the method is described correctly and in detail in [24], there is a misstatement of one of the assumptions, which is corrected in [26, §3.6] and briefly in Note 2 of the current chapter. Specifically, controlling an unmatched nominal covariate by stratifying matched pairs, as in Table 22.1 of Sect. 22.2, [24] and [26, §3.6], requires that, on the logit scale, the propensity score is the sum of two pieces, an arbitrary, typically unknown, function of the covariates controlled by matching plus another arbitrary, typically unknown, function of the nominal covariate controlled by stratifying pairs; see [26, Expression (3.13)]. Under this assumption, matching eliminates bias from the covariates controlled by matching, and conditioning on the observed imbalance in the unmatched but stratified covariate eliminates bias from that covariate [26, §3.6]. The method of Sect. 22.2 is the simplest case of a general method discussed in [23, 25]. Related methods are developed in [5, 18, 20].

Section 22.3 is derived from a paper with Jeffrey Silber [32] that contains further discussion and additional references. John Gerring and Jason Seawright [15, 36] and Sidney Tarrow [43] discuss the selection of informative cases in qualitative investigations. The literature on thick description, ethnography, and qualitative research is vast. The books by Lonnie Athens [1], Charles Bosk [4], and Sue

Estroff [12] are fine examples of ethnography touching upon criminology, surgery, and psychiatry, and the volume containing Becker's essay [2] contains many other interesting essays as well.

22.5 Software

In `R`, Ruoqi Yu's package `thickmatch` builds a large optimal match containing a small number of exceptionally close matched pairs for thick description. The package uses a technique briefly described in [31, §5.1].

References

1. Athens, L.: Violent Criminal Acts and Actors Revisited. University of Illinois Press, Urbana, IL (1997)
2. Becker, H.S.: The epistemology of qualitative research. In: Jessor, R., Colby, A., Shweder, R. (eds.) Ethnography and Human Development, pp. 53–72. University of Chicago Press, Chicago (1996)
3. Bittner, E., Garfinkel, H.: 'Good' organizational reasons for 'bad' organizational records. In: Garfinkel, H. (ed.) Studies in Ethnomethodology, pp. 186–207. Prentice Hall, Englewood Cliffs, NJ (1967)
4. Bosk, C.L.: Forgive and Remember: Managing Medical Failure. University of Chicago Press, Chicago (1981)
5. Branson, Z., Miratrix, L.W.: Randomization tests that condition on non-categorical covariate imbalance. J. Causal Infer. **7**, 1–29 (2019)
6. Cochran, W.G.: Analysis of covariance: its nature and uses. Biometrics **13**, 261–281 (1957)
7. Coleman, J., Hoffer, T., Kilgore, S.: Cognitive outcomes in public and private schools. Soc. Educ. **55**, 65–76 (1982)
8. Cox, D.R.: A note on data-splitting for the evaluation of significance levels. Biometrika **62**, 441–444 (1975)
9. Daniel, S., Armstrong, K., Silber, J.H., Rosenbaum, P.R.: An algorithm for optimal tapered matching, with application to disparities in survival. J. Comput. Graph. Stat. **17**, 914–924 (2008)
10. Erasmus: The Collected Works of Erasmus, vol. 34, Adages IIvii1 to IIIiii100. University of Toronto Press, Toronto (1992)
11. Ertefaie, A., Small, D.S., Rosenbaum, P.R.: Quantitative evaluation of the trade-off of strengthened instruments and sample size in observational studies. J. Am. Stat. Assoc. **113**, 1122–1134 (2018)
12. Estroff, S.E.: Making It Crazy: An Ethnography of Psychiatric Clients in an American Community. University of California Press, Berkeley (1985)
13. Geertz, C.: Thick description: toward an interpretative theory of culture. In: Geertz, C. (ed.) The Interpretation of Cultures, pp. 3–30. Basic Books , New York (1973)
14. Geertz, C.: Local Knowledge. Basic Books, New York (1983)
15. Gerring, J: Is there a (viable) crucial-case method. Comp. Polit. Stud. **40**, 231–253 (2007)
16. Goldberger, A.S., Cain, G.S.: The causal analysis of cognitive outcomes in the Coleman, Hoffer and Kilgore report. Soc. Educ. **55**, 103–122 (1982)
17. Heller, R., Rosenbaum, P.R., Small, D.: Split samples and design sensitivity in observational studies. J. Am. Stat. Assoc. **104**, 1090–1101 (2009)

18. Hennessy, J., Dasgupta, T., Miratrix, L., Pattanayak, C., Sarkar, P.: A conditional randomization test to account for covariate imbalance in randomized experiments. J. Causal Inference **4**, 61–80 (2016)
19. Karmakar, B., Small, D.S., Rosenbaum, P.R.: Using approximation algorithms to build evidence factors and related designs for observational studies. J. Comp. Graph. Stat. **28**, 698–709 (2019)
20. Miratrix, L.W., Sekhon, J.S., Yu, B.: Adjusting treatment effect estimates by post-stratification in randomized experiments. J. R. Stat. Soc. B **75**, 369–396 (2013)
21. Pimentel, S.D., Small, D.S., Rosenbaum, P.R.: Constructed second control groups and attenuation of unmeasured biases. J. Am. Stat. Assoc. **111**, 1157–1167 (2016)
22. Rosenbaum, P.R.: The consequences of adjustment for a concomitant variable that has been affected by the treatment. J. R. Stat. Soc. A **147**, 656–666 (1984)
23. Rosenbaum, P.R.: Conditional permutation tests and the propensity score in observational studies. J. Am. Stat. Assoc. **79**, 565–574 (1984)
24. Rosenbaum, P.R.: Permutation tests for matched pairs with adjustments for covariates. Appl. Stat. **37**, 401–411 (1988) (Correction: [26, §3.6.2, Expression (3.13)])
25. Rosenbaum, P.R.: Covariance adjustment in randomized experiments and observational studies (with Discussion). Stat. Sci. **17**, 286–327 (2002)
26. Rosenbaum, P.R.: Observational Studies, 2nd edn. Springer, New York (2002)
27. Rosenbaum, P.R.: Design sensitivity in observational studies. Biometrika **91**, 153–164 (2004)
28. Rosenbaum, P.R.: Heterogeneity and causality: unit heterogeneity and design sensitivity in observational studies. Am. Stat. **59**, 147–152 (2005)
29. Rosenbaum, P.R.: What aspects of the design of an observational study affect its sensitivity to bias from covariates that were not observed? In: Dorans, N.J., Sinharay, S. (eds.) Looking Back: Proceedings of a Conference in Honor of Paul W. Holland. Lecture Notes in Statistics Book, vol. 202, pp. 87–114. Springer, New York (2011)
30. Rosenbaum, P.R.: Impact of multiple matched controls on design sensitivity in observational studies. Biometrics **69**, 118–127 (2013)
31. Rosenbaum, P.R.: Imposing minimax and quantile constraints on optimal matching in observational studies. J. Comp. Graph. Stat. **26**, 66–78 (2017)
32. Rosenbaum, P.R., Silber, J.H.: Matching and thick description in an observational study of mortality after surgery. Biostatistics **2**, 217–232 (2001)
33. Rosenbaum, P.R., Silber, J.H.: Sensitivity analysis for equivalence and difference in an observational study of neonatal intensive care units. J. Am. Stat. Assoc. **104**, 501–511 (2009)
34. Rosenbaum, P.R., Silber, J.H.: Using the exterior match to compare two entwined matched comparison groups. Am. Stat. **67**, 67–75 (2013)
35. Ryle, G.: Collected Papers, vol. 2. Hutchinson, London (1971)
36. Seawright, J., Gerring, J.: Case selection techniques in case study research. Polit. Res. Quart. **61**, 294–308 (2008)
37. Silber, J.H., Rosenbaum, P.R., Trudeau, M.E., Even-Shoshan, O., Chen, W., Zhang, X., Mosher, R.E.: Multivariate matching and bias reduction in the surgical outcomes study. Med. Care **39**, 1048–1064 (2001)
38. Silber, J.H., Rosenbaum, P. R., Trudeau, M.E., Chen, W., Zhang, X., Lorch, S.L., Rapaport-Kelz, R., Mosher, R.E, Even-Shoshan, O.: Preoperative antibiotics and mortality in the elderly. Ann. Surg. **242**, 107–114 (2005)
39. Silber, J.H., Rosenbaum, P.R., Zhang, X., Even-Shoshan, O.: Estimating anesthesia and surgical time from medicare anesthesia claims. Anesthesiology **106**, 346–355 (2007)
40. Silber, J.H., Lorch, S.L., Rosenbaum, P.R., Medoff-Cooper, B., Bakewell-Sachs, S., Millman, A., Mi, L., Even-Shoshan, O., Escobar, G.E.: Additional maturity at discharge and subsequent health care costs. Health Serv. Res. **44**, 444–463 (2009)
41. Small, D.S., Rosenbaum, P.R.: War and wages: the strength of instrumental variables and their sensitivity to unobserved biases. J. Am. Stat. Assoc. **103**, 924–933 (2008)
42. Stone, M.: Cross-validatory choice and assessment of statistical predictions. J. R. Stat. Soc. B **36**, 111–147 (1974)

43. Tarrow, S.:The strategy of paired comparison: toward a theory of practice. Comp. Polit. Stud. **43**, 230–259 (2010)
44. Tukey, J.W.: Exploratory Data Analysis. Addison-Wesley, Reading, MA (1977)
45. Vandenbroucke, J.: In defence of case reports and case series. Ann. Intern. Med. **134**, 330–334 (2001)
46. Zhang, K., Small, D.S., Lorch, S., Srinivas, S., Rosenbaum, P.R.: Using split samples and evidence factors in an observational study of neonatal outcomes. J. Am. Stat. Assoc. **106**, 511–524 (2011)
47. Zhao, Q., Small, D.S., Rosenbaum, P.R.: Cross-screening in observational studies that test many hypotheses. J. Am. Stat. Assoc. **113**, 1070–1084 (2018)

Chapter 23
Planning the Analysis

Abstract "Make your theories elaborate" in observational studies, argued R.A. Fisher, so that the many predictions of such a theory may disambiguate the association between treatment and outcome. How should one plan the analysis of an observational study to check the predictions of an elaborate theory?

23.1 Plans Enable

A randomized clinical trial follows a protocol that describes the design of the trial and a plan for its analysis. The plan for analysis will identify a primary endpoint or outcome, possibly secondary endpoints, describe the comparisons that will be made, and so forth. The plan for a clinical trial is reviewed by a funding agency and, for large important trials, the plan may be published before the trial begins. In this sense, the plan is a public plan. There is no barrier to planning the analysis of an observational study, and there is much to gain from planning.

Plans enable. In all aspects of life, much can be done with a plan that cannot be done without one.

John Tukey [43] wrote:

> Important questions can demand the most careful planning for confirmatory analysis. ... Preplan the main analysis (having even two main analyses may be too many!) ... I see no real alternative, in most truly confirmatory studies, to having a single main question—in which a question is specified by ALL of design, collection, monitoring, and analysis.

A study that follows a public plan is more convincing than a study that emerges ambiguously from a fog of data. A public plan is subject to public scrutiny before the study begins, so criticism that would otherwise surface after the fact arises before the fact, perhaps leading to a better study more resistant to legitimate criticism. If there is a public plan, and if the critic raises no objection to the plan, then criticism that could have been raised to the plan rings hollow when raised for the first time after a particular conclusion is announced.

A planned analysis is more transparent than an unplanned analysis. As discussed in Chap. 6, transparency means making evidence evident. If there is a public plan for analysis, then the investigators either followed their plan or they deviated from

© Springer Nature Switzerland AG 2020

P. R. Rosenbaum, *Design of Observational Studies*, Springer Series in Statistics,
https://doi.org/10.1007/978-3-030-46405-9_23

it. If they deviated from the plan, they will be obligated to explain and justify the deviation. Perhaps the deviation will be judged appropriate, perhaps more appropriate in light of events than adherence to the plan. No matter. What was done and why it was done are open to view.

Without a plan, an analysis will first use the methods that are most familiar, most ready to hand, then suffer a disorganized retreat to appropriate methods when unforeseen but foreseeable dilemmas arise. Even if appropriate methods are ultimately used, such an analysis looks like what it is: a dubious circular meander around data, methods, and conclusions. Here is a common example. We are all more familiar with tests and confidence intervals for differences than with tests and intervals for near equivalence, although there is sound, straightforward, standard methodology available for both tasks [2, 6, 16]. Meandering through a data set, we therefore apply tests for difference indiscriminately, then remember that tests for difference make a muddle of attempts to demonstrate near equivalence, then retreat to appropriate methods in a manner that rarely convinces us, much less anyone else. And yet a moment's thought before the fact would have revealed that to provide evidence for a certain conclusion means finding a difference here and near equivalence there. A moment's thought before the fact is readily available; it just takes a moment's thought. This dilemma is not in the data but in the absence of a plan.[1]

A plan does not preclude unplanned analyses. A plan distinguishes planned and unplanned analyses.[2]

Admittedly, you cannot plan an ornate and delicate analysis. You can plan the analysis of a randomized trial because randomization has done the heavy lifting of ensuring that comparable people are compared; ornate analysis is not needed. You can plan the analysis of an observational study if matching in the design has done the heavy lifting of ensuring that ostensibly comparable people are compared; ornate adjustments of outcomes for observed covariates are not essential.[3] Properly performed, matching is done without examining outcomes; it is, therefore, part of design. To say that you cannot plan an ornate and delicate analysis is to offer a telling argument against ornate and delicate analyses.

Novelty and originality may be acceptable in a plan, but soundness, not novelty or originality, is the mark of excellence in a planned analysis. The analysis of a standard design may follow a standard plan.

[1] See, for instance, Footnote 2 in Chap. 5, where I deliberately did the meandering analysis I intend to correct in the current chapter.

[2] John Tukey: "We do not dare either give up exploratory data analysis or make it our sole interest" [44, page 72]. "We often forget how science and engineering function. Ideas come from previous exploration more often than from lightning strokes. Important questions can demand the most careful planning for confirmatory analysis" [43, page 23].

[3] By "ostensibly comparable," I mean comparable in terms of measured covariates, $\mathbf{x}$, leaving open the question of whether people who look comparable in terms of $\mathbf{x}$ are actually comparable in terms of unmeasured covariates u. "Ostensibly comparable" is an abbreviation.

23.2 Elaborate Theories

23.2.1 R.A. Fisher's Advice

In his paper, "The planning of observational studies of human populations," William Cochran [11, §5] wrote:

> About 20 years ago, when asked in a meeting what can be done in observational studies to clarify the step from association to causation, Sir Ronald Fisher replied: 'Make your theories elaborate.' The reply puzzled me at first, since by Occam's razor, the advice usually given is to make theories as simple as is consistent with known data. What Sir Ronald meant, as subsequent discussion showed, was that when constructing a causal hypothesis one should envisage as many different consequences of its truth as possible, and plan observational studies to discover whether each of these consequences is found to hold.... [T]his multiphasic attack is one of the most potent weapons in observational studies.

Fisher's advice fits closely with the devices in Sect. 5.2 that disambiguate causal associations [33], and with the reduction in sensitivity to unmeasured biases that may follow from an anticipated pattern of treatment effects in Chap. 18. In Fisher's advice, the temporal order is important: one envisions consequences of a causal hypothesis and then plans the study to discover whether the consequences hold.

A simple example discussed in Sects. 5.2.2, 12.3 and Chap. 21 is the use of two control groups. If the treatment is the cause of its ostensible effects, the anticipated pattern with two control groups is clear: the treated group differs substantially from the two control groups, which do not differ substantially from each other. The prediction here is for two substantial differences and one near equivalence. What plan is appropriate for the analysis?

A second simple example discussed in Sects. 5.2.3 and 18.2 concerns coherence of two outcomes, for instance, an increase in both math and verbal test scores associated with smaller class sizes prompted by Maimonides' rule. There is the hypothesis of a coherent association and hypotheses about two outcomes. What plan is appropriate for the analysis?

23.2.2 What Should a Planned Analysis Accomplish?

An elaborate theory makes an elaborate prediction, and in Chap. 18 it was seen that certain forms of compatibility with an elaborate theory may make the study less sensitive to unmeasured biases. The investigator hopes that the use of two control groups rather than one, or the search for coherence among two outcomes, will enlarge what can legitimately be said, not diminish it, but there are pitfalls for the unwary.

In the first instance, the data may confirm part of the elaborate theory, refute another part, with a third part left in an ambiguous state. A binary "confirm" or "otherwise" is not appropriate. The evaluation of an elaborate theory will, therefore,

involve more than one comparison. If several comparisons are made using statistical procedures, then each comparison runs a risk of error, with the resulting possibility that a series of small risks of error accumulate to a risk of error that is no longer small. If a correction for multiple testing were made to control this accumulation of error, say using the Bonferroni inequality, then the investigator might discover, to her shock and dismay, that if she had used the first control group alone, it would have had outcomes significantly different from the treated group, and if she had used the second control group alone, it too would have had outcomes significantly different from the treated group, but because she used both control groups and did two tests, neither difference is significant after correcting for multiple testing. This, too, is not appropriate.

Moreover, if the elaborate theory makes predictions of near equivalence, and if tests for difference are performed where tests for near equivalence are needed, then successes may be misclassified as failures and failures as successes. In a large study, two control groups might have nearly equivalent outcomes, yet differ significantly; this is a success that might be mistaken for a failure. Conversely, two control groups may have outcomes that do not differ significantly, and yet because of low power may provide no evidence that the outcomes are similar in the two groups; this is a failure that might be mistaken for a success. A test for near equivalence is one that can reject substantial inequivalence.

So a plan for analysis must do several things. To be a legitimate statistical analysis at all, it must control, in some appropriate sense, the risk of incorrect inferences. It must do this while making several comparisons, resulting in a graduated assessment of compatibility with the elaborate theory, rather than a simple "yes" or "no." If some comparisons have priority over others, then the analysis should respect that priority: the most important comparisons should come first, in some sense, and not be hindered by added interest in other comparisons, as seemed to be the case above in connection with the Bonferroni inequality.

To be credible in their respective roles, an elaborate theory and priorities for its assessment must appear in a plan that precedes examination of outcomes. The elaborate theory is intended to make predictions, but it is not making predictions if it is built to fit the data that will be used to assess those predictions. Priorities can aid in controlling the risk of incorrect inferences only if they shape the analysis rather than being shaped by it.

23.3 Three Simple Plans with Two Control Groups

23.3.1 The Simplest Plan for Analysis with Two Control Groups

To examine the logic of the simplest plan in a practical case, consider again Table 5.1 from the study by Bilban and Jakopin [7] of genetic damage among lead-zinc miners

compared with two control groups. This simplest plan is, in fact, too simple even for this most basic situation. Better plans will be considered soon. The focus, for the moment, is on the logic of the plan, not yet on finding a good plan.

Let us imagine that Bilban and Jakopin originally planned to use one control group, say the Slovene residents far from the mine. Let us imagine further that they added the second control group of local residents near the mine who were not miners in response to concerns that pollution in the area of the mine, not the mine itself, might be responsible for greater genetic damage among miners. This second control group is a reasonable response to that specific concern. Finally, let us suppose that they wanted to give priority to the first of these two control groups.

Consider the following analytic plan. Test the hypothesis of equality of micronuclei frequencies among miners and the first control group, say at level 0.05. If that hypothesis is not rejected, stop. If that hypothesis is rejected, declare it rejected and test the second hypothesis of equality of micronuclei frequencies among miners and the second control group, also at level 0.05, declaring it rejected or not as appropriate.

Admittedly, as analytic plans go, this is not much of a plan. It does nothing to investigate the magnitude of the effect, which is surely important, and it does nothing to appraise the comparability of the two control groups, which is also important. The issue of sensitivity to unmeasured biases is not mentioned. This plan, such as it is, does not plan much. Perhaps what it plans is less important than what it ignores. No matter. The plan exemplifies a certain principle in the simplest possible setting.

In Table 5.1 from [7], the first test yields a t-statistic of 16.13, the second a t-statistic of 13.87. If the distributions of micronuclei were independent and Normal with expectations μ_t, μ_{c1}, and μ_{c2} and standard deviations ω_t, ω_{c1}, and ω_{c2} in the treated group and the two control groups, with $c1$ referring to Slovene residents and $c2$ referring to local residents, then, at the 0.05 level, the plan would reject the first hypothesis, then test and reject the second.[4]

Notice first that, under this plan, the addition of a second control group costs the investigators nothing. If the hypothesis of no difference between miners and the first control group would have been rejected without the second control group, it would also be rejected by this plan in the presence of the second control group. So the situation is different from correcting for multiple testing using the Bonferroni inequality as described above. The second control group may allow more to be said, but less will not be said.

[4]In Table 5.1, it appears that ω_t, ω_{c1}, and ω_{c2} are not equal, so a pooled standard deviation is not used, and the "t-statistics" are not actually distributed with a t-distribution. Because the degrees of freedom are fairly large, the separate standard errors are used, and the "t-statistics" are compared with the standard Normal distribution, both here and in all calculations from Table 5.1. Although this is an imperfect approach, the largest imperfection in my analysis is in assuming the data are Normal merely because only means and standard deviations were reported. This example is being used to illustrate certain concepts while minimizing incidentals unrelated to those concepts.

The second thing to notice is that the plan controls the probability of at least one false rejection in two tests provided the two separate tests do their separate jobs of controlling the probability of a false rejection. It is important to understand what is being claimed, then why the claim is true. The claim is that if each test has level 0.05, falsely rejecting one time in 20, then the two-step plan tests two hypotheses but makes at least one false rejection with probability at most 0.05. It is like buying two lottery tickets rather than one without increasing your chance of winning, though not quite like that. Something has prevented an accumulation of risk of false rejection. What is that something? Consider the logic.

There are two null hypotheses, H_0 and $\widetilde{H}_0$. To be specific, the one-sided null hypotheses are $H_0 : \mu_t \leq \mu_{c1}$ and $\widetilde{H}_0 : \mu_t \leq \mu_{c2}$. Here, H_0 refers to the first control group and $\widetilde{H}_0$ refers to the second, but it does not really matter what they refer to; they could be any two hypotheses about anything. The plan is to test H_0, and only if H_0 is rejected to test $\widetilde{H}_0$. Consider the cases, one at a time. If H_0 and $\widetilde{H}_0$ are both false, then you cannot falsely reject anything, so there is no risk of false rejection; any rejection is a correct rejection. If H_0 is false but $\widetilde{H}_0$ is true, then you cannot falsely reject H_0, so to falsely reject anything you have to, among other things, falsely reject $\widetilde{H}_0$, but the chance of this is at most 0.05. If H_0 and $\widetilde{H}_0$ are both true, then you commit a false rejection if and only if you reject H_0, and the chance of this is at most 0.05. So no matter which case is true, the chance of falsely rejecting at least one true hypothesis is at most 0.05.[5]

This first bit of logic ramifies in many useful directions [3, 15, 16, 20, 22, 36]. This logic is discussed explicitly in the context of clinical trials by Gary Koch and S.A. Gansky [22] and Peter Bauer [3], but appears to have been understood 20 years earlier by Ruth Marcus et al. [28], who were doing something related but subtler; see also the unifying article by Jelle Goeman and Aldo Solari [15] The logic also has links to multiparameter tests [5, 23, 25] and has been combined in an interesting way with the Bonferroni inequality [19, 46, §3].

There is a second piece to the logic.

[5]Frank Yoon and colleagues [47, 48] improve upon this plan in a sensible way. Suppose that we can specify a "meaningful difference" (sometimes called a "clinically significant difference") of $\tau_{\text{mf}} \geq 0$. The notion is that a difference smaller than τ_{mf} may exist but be too small to merit attention. Yoon suggests testing $H_0^{(\tau)} : \mu_t - \tau \leq \mu_{c1}$ in order [36] at level α for $\tau \in (-\infty, \tau_{\text{mf}}]$, stopping with the smallest $H_0^{(\tau)}$ that is not rejected, say τ_{nr}, declaring with $1 - \alpha$ confidence that $\mu_t - \mu_{c1} \geq \tau_{\text{nr}}$; however, if all $\tau \in (-\infty, \tau_{\text{mf}}]$ are rejected, declaring with $1 - \alpha$ confidence that $\mu_t - \mu_{c1} \geq \tau_{\text{mf}}$ and continuing to test $\widetilde{H}_0^{(\tau)} : \mu_t - \tau \leq \mu_{c2}$, in order at level α for $\tau \in (-\infty, \tau_{\text{mf}}]$, with a parallel interpretation. His method tries to establish a meaningful difference for the first control group and, if that is successful, tries to establish a meaningful difference for the second control group.

23.3.2 A Symmetric Plan for Analysis with Two Control Groups

An awkward feature of the first plan is that the investigators had to give priority to one of the control groups. Can the two control groups be viewed symmetrically?

Consider a second plan. Test at level 0.05 the hypothesis that $\overline{H}_0 : \mu_t \leq (\mu_{c1} + \mu_{c2})/2$, which says that the expected level of micronuclei in the treated group is no higher than in the average of the two control groups. If $\overline{H}_0$ is not rejected, stop. If $\overline{H}_0$ is rejected, declare it rejected and test at the 0.05 level both $H_0 : \mu_t \leq \mu_{c1}$ and $\widetilde{H}_0 : \mu_t \leq \mu_{c2}$, rejecting neither, either, or both as appropriate.

This second plan is little better than the first, but it does exemplify a second piece of the logic in the simplest case. Here, the two control groups are handled symmetrically, and if the control groups are similar, there is likely to be more power in testing $\overline{H}_0$ than in testing either H_0 or $\widetilde{H}_0$, because in testing $\overline{H}_0$ all the controls are used at once.

If each of the tests is done with level 0.05, then the chance of at least one false rejection is at most 0.05, even though three tests may be performed. To see this, consider the possible cases. If $\overline{H}_0$ is true, then you falsely reject at least one true hypothesis if and only if you reject $\overline{H}_0$, and the chance of this is at most 0.05. Now suppose $\overline{H}_0$ is false. Then you cannot falsely reject $\overline{H}_0$; any rejection of $\overline{H}_0$ is a correct rejection. Moreover, if $\overline{H}_0$ is false, then $\mu_t > (\mu_{c1} + \mu_{c2})/2$, so $H_0 : \mu_t \leq \mu_{c1}$ and $\widetilde{H}_0 : \mu_t \leq \mu_{c2}$ cannot both be true; either H_0 or $\widetilde{H}_0$, or both, are false. If both H_0 and $\widetilde{H}_0$ are false, then a false rejection is impossible. If exactly one of H_0 and $\widetilde{H}_0$ is true, then to falsely reject at least one hypothesis you must falsely reject that one true hypothesis, and the chance of this is at most 0.05.

In Table 5.1 from the study by Bilban and Jakopin [7], the hypothesis $\overline{H}_0$ is tested by estimating the contrast $1 \times \mu_t - \left(\frac{1}{2}\right) \times \mu_{c1} - \left(\frac{1}{2}\right) \times \mu_{c2}$ as 8.25 using the sample means in place of the population means, with estimated standard error

$$0.519 = \sqrt{\left(1^2\right) \cdot 0.479 + \left(-\frac{1}{2}\right)^2 \cdot 0.143 + \left(-\frac{1}{2}\right)^2 \cdot 0.377}, \tag{23.1}$$

yielding a "t-statistic" of $8.25/0.519 = 15.9$, so $\overline{H}_0$ is rejected, and both H_0 and $\widetilde{H}_0$ are tested, with t-statistics of 16.1 and 13.9 as in Table 5.1, so all three hypotheses are rejected.

Where the first plan added a second control group at no cost, the second plan may confer a gain, because the sample size is increased in the first step by using both control groups at once. Perhaps $\overline{H}_0$ will be rejected when neither H_0 nor $\widetilde{H}_0$ is rejected. If the treated group is judged to produce responses that, on average, are higher than the two control groups, then the investigation continues by testing the two control groups one at a time.

The first bit of logic concerned testing hypotheses in an order of priority. The second bit of logic, which was just illustrated, concerned the possibility that some

hypotheses preclude one another, so in testing several hypotheses you may know that they cannot all be true simultaneously, even though you do not know which one is true. The second bit of logic is used in building confidence intervals: then you test infinitely many hypotheses, but only one hypothesis is true, so no correction for multiple testing is needed. The second bit of logic shows up in a certain form in two methods for testing several hypotheses, specifically the method of Ruth Marcus et al. [28] and the method of Julliet Popper Shaffer [40], together with a large literature developing from these two papers. In the specific form used above to compare two control groups, H_0 and $\widetilde{H}_0$ do not preclude one another, but they do preclude one another if $\overline{H}_0$ is false, and in this case $\langle \overline{H}_0, \{H_0, \widetilde{H}_0\}\rangle$ is called a "sequentially exclusive partition" of the three hypotheses, $\overline{H}_0$, H_0, $\widetilde{H}_0$, in the sense that a set of hypotheses in the sequence, here $\{H_0, \widetilde{H}_0\}$, contains at most one true hypothesis if all previous hypotheses in the sequence, here $\overline{H}_0$, are false; see [36].

Before developing a serious plan for an analysis with two control groups, one more bit of logic is needed, this time about equivalence tests.

23.3.3 Are the Two Control Groups Nearly Equivalent?

The current section is concerned with demonstrating that the two control groups are nearly equivalent in their responses, that is, with an equivalence test [6]. To assert that the two control groups are close in their responses is to assert that $|\mu_{c1} - \mu_{c2}|$ is small, say $|\mu_{c1} - \mu_{c2}| < \delta$, for some specified $\delta > 0$. To reject the hypothesis that $H_{\neq}^{(\delta)} : |\mu_{c1} - \mu_{c2}| \geq \delta$ is to provide a basis for asserting $|\mu_{c1} - \mu_{c2}| < \delta$; that is, to reject the hypothesis $H_{\neq}^{(\delta)}$ of inequivalence is to be in a position to assert equivalence, $|\mu_{c1} - \mu_{c2}| < \delta$. Moreover, $H_{\neq}^{(\delta)} : |\mu_{c1} - \mu_{c2}| \geq \delta$ is true if either $\overleftarrow{H}_0^{(\delta)} : \mu_{c1} - \mu_{c2} \leq -\delta$ or $\overrightarrow{H}_0^{(\delta)} : \mu_{c1} - \mu_{c2} \geq \delta$ is true; that is, $H_{\neq}^{(\delta)}$ is the union of two hypotheses [5, 25], and the set comprised of these two hypotheses, $\left\{\overleftarrow{H}_0^{(\delta)}, \overrightarrow{H}_0^{(\delta)}\right\}$, is exclusive in the sense that at most one of the two hypotheses is true [36].

Suppose that we adopt the plan of testing both $\overleftarrow{H}_0^{(\delta)}$ and $\overrightarrow{H}_0^{(\delta)}$ in one-sided, 0.05 level tests, rejecting each hypothesis if the corresponding P-value is less than 0.05, rejecting $H_{\neq}^{(\delta)}$ if both $\overleftarrow{H}_0^{(\delta)}$ and $\overrightarrow{H}_0^{(\delta)}$ are rejected. In brief, two tests are performed, each at level 0.05, with four possible outcomes: (1) no rejections, (2) reject $\overleftarrow{H}_0^{(\delta)}$ but not $\overrightarrow{H}_0^{(\delta)}$, (3) reject $\overrightarrow{H}_0^{(\delta)}$ but not $\overleftarrow{H}_0^{(\delta)}$, (4) reject $\overleftarrow{H}_0^{(\delta)}$, $\overrightarrow{H}_0^{(\delta)}$, and $H_{\neq}^{(\delta)}$. What is the chance of falsely rejecting at least one true hypothesis? The chance is at most 0.05. To see this, recall that $\left\{\overleftarrow{H}_0^{(\delta)}, \overrightarrow{H}_0^{(\delta)}\right\}$ is exclusive: at most one of $\overleftarrow{H}_0^{(\delta)}$ and $\overrightarrow{H}_0^{(\delta)}$ is true. If neither $\overleftarrow{H}_0^{(\delta)}$ nor $\overrightarrow{H}_0^{(\delta)}$ is true, then there is no true hypothesis to falsely reject, and the chance of a false rejection is zero. If $\overleftarrow{H}_0^{(\delta)}$ is true, then $\overrightarrow{H}_0^{(\delta)}$ is false, and a false rejection occurs if and only if $\overleftarrow{H}_0^{(\delta)}$ is rejected, which happens with

probability at most 0.05. If $\overrightarrow{H}_0^{(\delta)}$ is true, then $\overleftarrow{H}_0^{(\delta)}$ is false, and a false rejection occurs if and only if $\overrightarrow{H}_0^{(\delta)}$ is rejected, which happens with probability at most 0.05. With $\delta = 2$ in Table 5.1 from [7], two "t-statistics" are computed as

$$\overleftarrow{t} = \frac{(6.400 - 6.005) + 2}{\sqrt{0.143^2 + 0.377^2}} = 5.9 \text{ and } \overrightarrow{t} = \frac{(6.400 - 6.005) - 2}{\sqrt{0.143^2 + 0.377^2}} = -4.0, \tag{23.2}$$

where $\overleftarrow{H}_0^{(\delta)} : \mu_{c1} - \mu_{c2} \leq -\delta$ is rejected for large values of $\overleftarrow{t}$ and $\overrightarrow{H}_0^{(\delta)} : \mu_{c1} - \mu_{c2} \geq \delta$ is rejected for small values of $\overrightarrow{t}$. Here, both $\overleftarrow{t} \geq 1.65$ and $\overrightarrow{t} \leq -1.65$, where $\Phi(-1.65) = 0.05$ and $1 - \Phi(1.65) = 0.05$, give the approximate one-sided critical values from the Normal distribution. So we are in a position to assert that $|\mu_{c1} - \mu_{c2}| < \delta = 2$ having rejected the two alternatives to this assertion.

How should one define equivalence? That is, how should one pick a δ? In fact, there is no need to pick a single δ. Start with $\delta = \infty$, which corresponds with $\overleftarrow{t} = \infty$ and $\overrightarrow{t} = -\infty$, continuing to test smaller values of δ until a value δ^* is encountered such that either $\overleftarrow{H}_0^{(\delta)}$ or $\overrightarrow{H}_0^{(\delta)}$ is not rejected; then assert with 95% confidence that $|\mu_{c1} - \mu_{c2}| \leq \delta^*$. In Table 5.1, $\delta^* = 1.06$, so with 95% confidence, the frequency of micronuclei in the two control groups differs by at most 1.06.

The reasoning here closely parallels the reasoning behind the TOST, or two-one-sided tests procedure for equivalence testing [6, 39, 45] and the associated confidence interval [2, 21].[6] It is also closely related to the intersection–union principle; see [5, 6].

23.3.4 An Elementary Plan for Analysis with Two Control Groups

With two control groups, an elementary plan for analysis follows. The plan is stated in terms of rejection at the conventional level $\alpha = 0.05$, but any level α may be used instead.

[6] Alternatively, if somewhat unconventionally, one may view this as testing hypotheses in order [36] with an infinite sequentially exclusive partition of hypotheses, $\left\{\overleftarrow{H}_0^{(\delta)}, \overrightarrow{H}_0^{(\delta)}\right\}$, $\delta \in (0, \infty)$, where $\left\{\overleftarrow{H}_0^{(\delta)}, \overrightarrow{H}_0^{(\delta)}\right\}$ is "before" $\left\{\overleftarrow{H}_0^{(\delta')}, \overrightarrow{H}_0^{(\delta')}\right\}$ if $\delta > \delta'$. Notice this ordering: extreme hypotheses are tested first. This slightly unconventional view has two advantages in the current discussion. First, it allows equivalence tests to be embedded in other forms of sequentially exclusive partitions without requiring a separate formalism. This will be done in a moment. Second, for a single δ, it permits rejection of either $\overleftarrow{H}_0^{(\delta)}$ or $\overrightarrow{H}_0^{(\delta)}$ when both hypotheses cannot be rejected, something that is not always explicitly stated in discussions of TOST.

Step 1: Test $\overline{H}_0 : \mu_t \leq (\mu_{c1} + \mu_{c2})/2$. If the P-value is greater than 0.05, stop; otherwise, reject $\overline{H}_0$ and perform Step 2.

Step 2: Test both $H_0 : \mu_t \leq \mu_{c1}$ and $\widetilde{H}_0 : \mu_t \leq \mu_{c2}$, rejecting H_0 if its P-value is at most 0.05, rejecting $\widetilde{H}_0$ if its P-value is at most 0.05. If either P-value is above 0.05, stop. If both P-values are at most 0.05, perform Step 3.

Step 3: Starting with $\delta = \infty$ and proceeding to smaller values of δ, test both $\overleftarrow{H}_0^{(\delta)} : \mu_{c1} - \mu_{c2} \leq -\delta$ and $\overrightarrow{H}_0^{(\delta)} : \mu_{c1} - \mu_{c2} \geq \delta$. If the P-value for $\overleftarrow{H}_0^{(\delta)}$ is at most 0.05, reject $\overleftarrow{H}_0^{(\delta)}$, and if the P-value for $\overrightarrow{H}_0^{(\delta)}$ is at most 0.05, reject $\overrightarrow{H}_0^{(\delta)}$. If either P-value is greater than 0.05, stop testing and assert with 95% confidence that $|\mu_{c1} - \mu_{c2}| \leq \delta$. If both P-values are at most 0.05, then continue testing smaller values of δ.

This three step plan may reach a variety of conclusions. It may decide that there is insufficient evidence to reject $\overline{H}_0$, so that even when combining the sample sizes from the two control groups, the treated group cannot be confidently asserted to have higher responses than the average of the two control groups. If $\overline{H}_0$ is rejected in Step 1, then with 95% confidence one may assert $\mu_t > (\mu_{c1} + \mu_{c2})/2$. In Step 2, neither, either, or both of H_0 and $\widetilde{H}_0$ may be rejected; that is, one may be confident that the treated group has higher responses than neither control group, than one control group but not the other, or than both control groups. If H_0 and $\widetilde{H}_0$ are both rejected in Step 2, then Step 3 provides a confidence statement for the degree of equivalence of the outcomes in the two control groups.

As we have already seen, when applied to Table 5.1 from [7], the three-step procedure rejects $\overline{H}_0$, rejects both H_0 and $\widetilde{H}_0$, and rejects both $\overleftarrow{H}_0^{(\delta)}$ and $\overrightarrow{H}_0^{(\delta)}$ for all $\delta > 1.06$. With 95% confidence, the procedure asserts that the treated responses are above the responses in both control groups and the two control group means differ by $|\mu_{c1} - \mu_{c2}| \leq 1.06$.

The chance that this three-step procedure tests and falsely rejects at least one true hypothesis is at most 0.05. Perhaps this is intuitive now, having previously considered the steps in isolation, one at a time. A proof of a general kind is given in the appendix to this chapter. It uses the simple fact that

$$\left\langle \overline{H}_0,\ \left\{H_0,\ \widetilde{H}_0\right\},\ \left\{\overleftarrow{H}_0^{(\delta)},\ \overrightarrow{H}_0^{(\delta)}\right\},\ \delta \in (0, \infty) \right\rangle \tag{23.3}$$

is a sequentially exclusive partition of hypotheses, with $\left\{\overleftarrow{H}_0^{(\delta)},\ \overrightarrow{H}_0^{(\delta)}\right\}$ placed before $\left\{\overleftarrow{H}_0^{(\delta')},\ \overrightarrow{H}_0^{(\delta')}\right\}$ whenever $\delta > \delta'$; that is, as one proceeds forward in this sequence of sets of hypotheses, at most one hypothesis in a set is true when all the hypotheses in earlier sets are false. This is true because if $\overline{H}_0$ is false, then at most one hypothesis in $\left\{H_0,\ \widetilde{H}_0\right\}$ is true, and for all δ, at most one hypothesis in $\left\{\overleftarrow{H}_0^{(\delta)},\ \overrightarrow{H}_0^{(\delta)}\right\}$ is true. Intuitively, in the three-step procedure, whenever two hypotheses are tested at once, at most one of them is true, so you are taking only one 0.05 chance, not two 0.05 chances, of a false rejection.

23.3.5 An Alternative Plan for Analysis with Two Control Groups

A weakness of the analysis plan just described is that it might reject $H_0 : \mu_t \leq \mu_{c1}$ and $\widetilde{H}_0 : \mu_t \leq \mu_{c2}$ in Step 2, even though the treated group is close to one of the control groups, in the sense that $\mu_t - \max(\mu_{c1}, \mu_{c2})$ is small, and the two control groups are far apart, in the sense that $\max(\mu_{c1}, \mu_{c2}) - \min(\mu_{c1}, \mu_{c2})$ is large. One might reasonably argue that a difference between the control groups is problematic unless it is a small difference compared with the difference between the treated group and both control groups. We are not so interested in whether the difference between the two control groups is significantly different from zero, but rather whether the magnitude of bias needed to explain any difference between the control groups is far smaller than the magnitude of bias needed to explain the difference between the treated group and both control groups. More precisely, for some fixed number, τ, one might wish to assert

$$(\mu_t - \tau) - \max(\mu_{c1}, \mu_{c2}) > \max(\mu_{c1}, \mu_{c2}) - \min(\mu_{c1}, \mu_{c2}) \tag{23.4}$$

and one would have grounds for asserting this if one had rejected the hypothesis

$$H_{\diamond}^{(\tau)} : (\mu_t - \tau) - \max(\mu_{c1}, \mu_{c2}) \leq \max(\mu_{c1}, \mu_{c2}) - \min(\mu_{c1}, \mu_{c2}) . \tag{23.5}$$

For instance, rejection at level 0.05 of $H_{\diamond}^{(\tau)}$ with $\tau = 0$ asserts with 95% confidence that the treated mean, μ_t, exceeds the means, μ_{c1} and μ_{c2}, in both control groups by more than the control groups differ from each other.

The following procedure from [36] tests $H_{\diamond}^{(\tau)}$.

Step 1a: Test $\overline{H}_0^{(\tau)} : (\mu_t - \tau) \leq (\mu_{c1} + \mu_{c2})/2$. If the P-value is greater than 0.05, stop; otherwise, reject $\overline{H}_0^{(\tau)}$ and perform Step 2.

Step 2a: Test both $H_0^{(\tau)} : (\mu_t - \tau) \leq \mu_{c1}$ and $\widetilde{H}_0^{(\tau)} : \mu_t - \tau \leq \mu_{c2}$, rejecting $H_0^{(\tau)}$ if its P-value is at most 0.05, rejecting $\widetilde{H}_0^{(\tau)}$ if its P-value is at most 0.05. If either P-value is above 0.05, stop. If both P-values are at most 0.05, perform Step 3.

Step 3a: Test both $H_{\triangleright}^{(\tau)} : (\mu_t - \tau) - \mu_{c1} \leq \mu_{c1} - \mu_{c2}$ and $H_{\triangleleft}^{(\tau)} : (\mu_t - \tau) - \mu_{c2} \leq \mu_{c2} - \mu_{c1}$, rejecting $H_{\triangleright}^{(\tau)}$ if its P-value is at most 0.05 and rejecting $H_{\triangleleft}^{(\tau)}$ if its P-value is at most 0.05. If either P-value is above 0.05, stop. Otherwise, if both P-values are at most 0.05, reject $H_{\diamond}^{(\tau)}$.

For $\tau = 0$, Steps 1a and 2a are the same as Steps 1 and 2, so the conclusions about $\overline{H}_0$, H_0, and $\widetilde{H}_0$ are the same. In Step 3a, two cases need to be distinguished, namely $\max(\mu_{c1}, \mu_{c2}) = \mu_{c1}$ and $\max(\mu_{c1}, \mu_{c2}) = \mu_{c2}$. If $\max(\mu_{c1}, \mu_{c2}) = \mu_{c1}$, then $H_{\diamond}^{(\tau)}$ and $H_{\triangleright}^{(\tau)}$ are the same. Conversely, if $\max(\mu_{c1}, \mu_{c2}) = \mu_{c2}$, then $H_{\diamond}^{(\tau)}$ and

$H_{\triangleleft}^{(\tau)}$ are the same. Therefore, $H_{\diamond}^{(\tau)}$ is false if both $H_{\triangleright}^{(\tau)}$ and $H_{\triangleleft}^{(\tau)}$ are false; so, in Step 3a, $H_{\diamond}^{(\tau)}$ is rejected if both $H_{\triangleright}^{(\tau)}$ and $H_{\triangleleft}^{(\tau)}$ are rejected.

In Table 5.1 from the study by Bilban and Jakopin [7], the five hypotheses in Steps 1a–3a are each tested by a contrast in group means. For instance, in Step 1a, $\overline{H}_0^{(\tau)}$ is the hypothesis that $1 \times \mu_t - \left(\frac{1}{2}\right) \times \mu_{c1} - \left(\frac{1}{2}\right) \times \mu_{c2} \leq \tau$, and in Step 3a, $H_{\triangleright}^{(\tau)}$ is the hypothesis that $1 \times \mu_t - 2 \times \mu_{c1} + 1 \times \mu_{c2} \leq \tau$. As before, for $\tau = 0$, in Step 1a, the "t-statistic" for $\overline{H}_0^{(\tau)}$ is 15.9, so $\overline{H}_0^{(\tau)}$ is rejected at the 0.05 level, and the procedure continues to Step 2a. Had $\overline{H}_0^{(\tau)}$ not been rejected, the procedure would have stopped in Step 1a. In Step 2a, the "t-statistics" for $H_0^{(\tau)}$ and $\widetilde{H}_0^{(\tau)}$ are, respectively, 16.1 and 13.8, so $H_0^{(\tau)}$ and $\widetilde{H}_0^{(\tau)}$ are both rejected at the 0.05 level and the procedure continues to Step 3a. Had either $H_0^{(\tau)}$ or $\widetilde{H}_0^{(\tau)}$ not been rejected, the procedure would have stopped in Step 2a. In Step 3a, the "t-statistics" for $H_{\triangleright}^{(\tau)}$ and $H_{\triangleleft}^{(\tau)}$ are, respectively 11.4 and 9.8, so both $H_{\triangleright}^{(\tau)}$ and $H_{\triangleleft}^{(\tau)}$ are rejected, and hence $H_{\diamond}^{(\tau)}$ is rejected. Therefore, we may assert with 95% confidence that the treated group mean exceeds both of the control group means by more than the control group means differ from each other. After considering the properties of this procedure, other values of τ will be evaluated.

In Steps 1a, 2a, and 3a, the chance of testing and rejecting at least one true hypothesis is at most 0.05. The reasoning here closely parallels what has been done before, except there are more steps of the same kind. The details of the reasoning constitute the remainder of this somewhat long, slightly technical paragraph, which you may skip if that is your preference. The general argument is given more concisely in the appendix to this chapter. First, we check that the sequence of hypotheses has the needed structure. The sequence of sets of hypotheses

$$\left\langle \overline{H}_0^{(\tau)},\ \left\{H_0^{(\tau)},\ \widetilde{H}_0^{(\tau)}\right\},\ \left\{H_{\triangleright}^{(\tau)},\ H_{\triangleleft}^{(\tau)}\right\}\right\rangle \tag{23.6}$$

is a sequentially exclusive partition in the sense that: (i) if $\overline{H}_0^{(\tau)}$ is false, then at most one hypothesis in $\left\{H_0^{(\tau)},\ \widetilde{H}_0^{(\tau)}\right\}$ is true, and (ii) if $\overline{H}_0^{(\tau)}$, $H_0^{(\tau)}$, $\widetilde{H}_0^{(\tau)}$ are all false, then at most one hypothesis in $\left\{H_{\triangleright}^{(\tau)},\ H_{\triangleleft}^{(\tau)}\right\}$ is true. To see this, note first that part (i) follows from the same reasoning used for $\overline{H}_0$, H_0, and $\widetilde{H}_0$. For part (ii), if $H_0^{(\tau)}$: $(\mu_t - \tau) \leq \mu_{c1}$ and $\widetilde{H}_0^{(\tau)} : \mu_t - \tau \leq \mu_{c2}$ are both false, then $(\mu_t - \tau) - \mu_{c1} > 0$ and $(\mu_t - \tau) - \mu_{c2} > 0$, but either $\mu_{c1} - \mu_{c2} \leq 0$ or $\mu_{c2} - \mu_{c1} \leq 0$, so at most one of the following two hypotheses is true $H_{\triangleright}^{(\tau)} : (\mu_t - \tau) - \mu_{c1} \leq \mu_{c1} - \mu_{c2}$ and $H_{\triangleleft}^{(\tau)} : (\mu_t - \tau) - \mu_{c2} \leq \mu_{c2} - \mu_{c1}$. Next, we verify the claim about the probability of at least one false rejection. There is, of course, a general pattern here, in which a sequentially exclusive partition of hypotheses, tested in an order such as Steps 1a–3a, will falsely reject a true hypothesis with probability at most 0.05. This is demonstrated in general in [36] and in the appendix to this chapter, but let us do

one last particular case, namely Steps 1a–3a. If $\overline{H}_0^{(\tau)}$ is true, then a false rejection occurs if and only if $\overline{H}_0^{(\tau)}$ is rejected, which happens with probability at most 0.05. So suppose $\overline{H}_0^{(\tau)}$ is false; then either $H_0^{(\tau)}$ or $\widetilde{H}_0^{(\tau)}$, or both, are false. If exactly one of $H_0^{(\tau)}$ or $\widetilde{H}_0^{(\tau)}$ is true, then at least one false rejection occurs if and only if that one true hypothesis is rejected, which happens with probability at most 0.05. If both $H_0^{(\tau)}$ and $\widetilde{H}_0^{(\tau)}$are false, then a false rejection cannot occur at Step 2a. So suppose $\overline{H}_0^{(\tau)}$, $H_0^{(\tau)}$, and $\widetilde{H}_0^{(\tau)}$are false; then either $H_{\rhd}^{(\tau)}$ or $H_{\lhd}^{(\tau)}$, or both, are false. If both are false, a false rejection is not possible. If exactly one of $H_{\rhd}^{(\tau)}$ or $H_{\lhd}^{(\tau)}$ is true, then a false rejection occurs if and only if that one true hypothesis is rejected, which happens with probability at most 0.05.

There is no need to select a value of τ. Rather, apply Steps 1a–3a repeatedly, beginning with $\tau = -\infty$, considering larger and larger values of τ, until for the first time at τ^* the procedure stops because one of the hypotheses has not been rejected. Then declare with 95% confidence that

$$\left(\mu_t - \tau^*\right) - \max\left(\mu_{c1},\ \mu_{c2}\right) > \max\left(\mu_{c1},\ \mu_{c2}\right) - \min\left(\mu_{c1},\ \mu_{c2}\right). \tag{23.7}$$

In Table 5.1, the first acceptance occurs at $\tau^* = 6.56$ when $H_{\rhd}^{(\tau)}$ is just barely not rejected in Step 3a with a "t-statistic" of 1.64. In words, with 95% confidence, the mean level of micronuclei in the treated group exceeds both control group means by at least 6.56 more than the two control groups differ from each other.

Another example is discussed in [36]. There, unlike Table 5.1, the two control groups differ significantly from each other, but with 95% confidence, the difference between the treated group and both control groups is substantially larger than the difference between the two control groups.

23.3.6 Summary

In this section, several plans have been considered for the analysis of a study with a treated group and two control groups. Some of the plans were too primitive for practical use and were simply illustrations of general principles. Other plans are suitable for a primary analysis. With two control groups, the "elaborate theory" predicts that the treated group will differ substantially from the control groups, which do not differ substantially from each other. The "elaborate theory" predicts a difference here and equivalence there, and the plan uses both tests for difference and for equivalence at appropriate steps. The plan tests the most important elements of the predictions first, without a correction for multiple testing; then, it continues to try to establish the less important elements of the prediction. The planned analysis may yield confirmation of some predictions and not others, and may provide quantitative statements that have some resemblance to confidence intervals.

The current section has considered the simplest of study designs without a sensitivity analysis. Sensitivity analyses with multiple control groups are possible, but they would take us a little far afield; see [34, §8]. The next section will plan a sensitivity analysis for coherence between two outcomes.

23.4 Sensitivity Analysis for Two Outcomes and Coherence

In Sects. 5.2.3 and 18.2, it was found that coherence among several outcomes has the potential to reduce sensitivity to departures from random treatment assignment. With two outcomes, what is a suitable plan for analysis?

In Table 5.3, Angrist and Lavy's study of the effects of class size on test performance was found to be somewhat less sensitive for the coherent combination of math and verbal test scores than for either score alone. The general results in Sect. 18.2 suggested that this is to be expected with two outcomes that are similarly affected by treatment but imperfectly correlated.

In our initial examination of Table 5.3, there was no plan for analysis. Three tests were done for two outcomes. Are there problems of multiple testing, of increased risk of false positives from performing several tests?

Suppose that there are two outcomes, say $\left(r_{Tij}, r_{Cij}\right)$ and $\left(\widetilde{r}_{Tij}, \widetilde{r}_{Cij}\right)$, with corresponding treated-minus-control matched pair differences Y_i and $\widetilde{Y}_i$, say the math and verbal test scores in Sect. 5.2.3. The null hypothesis of no treatment effect for the first outcome says $H_0 : r_{Tij} = r_{Cij}$, for all i, j, and the null hypothesis of no treatment effect for the second outcome is parallel, $\widetilde{H}_0 : \widetilde{r}_{Tij} = \widetilde{r}_{Cij}$, for all i, j. Write $\overline{H}_0$ for the hypothesis that is the conjunction of these two hypotheses, H_0 and $\widetilde{H}_0$, or in logical notation, $H_0 \wedge \widetilde{H}_0$, which says neither outcome was affected. Consider the following plan.

Step 1b: Test $\overline{H}_0$ using the coherent signed rank test. If the P-value is above 0.05, stop and do not reject $\overline{H}_0$. If the P-value is at most 0.05, reject $\overline{H}_0$ and perform Step 2b.

Step 2b: Test both H_0 and $\widetilde{H}_0$ using Wilcoxon's signed rank test, rejecting H_0 if its P-value is at most 0.05, rejecting $\widetilde{H}_0$ if its P-value is at most 0.05.

If the three separate P-values do their separate jobs of falsely rejecting a true hypothesis with probability at most 0.05, then the chance that the three-step procedure falsely rejects at least one true hypothesis is at most 0.05. The reasoning is parallel to that in Sect. 23.3. If $\overline{H}_0$ is false, it must be the case that either H_0 or $\widetilde{H}_0$, or both, are false. That is, $\left\langle \overline{H}_0,\ \left\{H_0,\ \widetilde{H}_0\right\}\right\rangle$ is a sequentially exclusive partition of hypotheses; if $\overline{H}_0$ is false, there is at most one true hypothesis in $\left\{H_0,\ \widetilde{H}_0\right\}$. If $\overline{H}_0$ is true, then at least one true hypothesis is rejected if and only if $\overline{H}_0$ is rejected in Step 1b, and this happens with probability at most 0.05. If $\overline{H}_0$ is false, then either H_0 or $\widetilde{H}_0$ is false, so at most one true hypothesis is tested in Steps 1b and 2b, and the chance of falsely rejecting that one true hypothesis is at most 0.05. See [24] for a closely related procedure.

The sensitivity analysis uses the upper bounds on the P-values in Steps 1b and 2b. The appendix to this chapter and [38] discusses why this procedure works.

In Table 5.3, the two-step procedure rejects $\overline{H}_0$, H_0, and $\widetilde{H}_0$ for $\Gamma \leq 1.4$. For $\Gamma = 1.55$, it rejects $\overline{H}_0$ and $\widetilde{H}_0$, but not H_0. For $\Gamma = 1.65$, it rejects $\overline{H}_0$ but not H_0 nor $\widetilde{H}_0$. In other words, the coherent association using both outcomes is marginally less sensitive to deviations from random assignment than either outcome alone. In this instance, the plan consisting of Steps 1b and 2b would have supported our previous interpretation of Table 5.3, while addressing any concern about testing more than one hypothesis. The general results in Sect. 18.2 provide some guidance about when $\overline{H}_0$ will be less sensitive to bias than H_0 or $\widetilde{H}_0$.

If instead a Bonferroni adjustment had been made for three tests, none of the hypotheses would have been rejected at $\Gamma = 1.55$. If the coherent test was not used, and Wilcoxon's statistic was applied twice, once to H_0 and once to $\widetilde{H}_0$, with a Bonferroni adjustment, then neither H_0 nor $\widetilde{H}_0$ would be rejected for $\Gamma \geq 1.45$.[7] For discussion of sensitivity analysis adjusted by the Bonferroni inequality, see [13, 17, 49] and [38, §4.5].

The test of the coherent null hypothesis—that is, the first hypothesis, $\overline{H}_0$, in the sequence—need not use the coherent signed rank statistic with its fixed, equal weights. For instance, $\overline{H}_0$ might be tested using a hedged bet on equal weights, as discussed in Sect. 18.2 and [37]. With such a hedged bet, $\overline{H}_0$ might be rejected, and testing might continue on to test both H_0 and $\widetilde{H}_0$ each at level 0.05, even if equal weights are a poor choice of weights.

23.5 Sensitivity Analysis for Tests of Equivalence

The absence of a treatment effect may be an important finding. See, for instance, the book *Costs, Risks and Benefits of Surgery* edited by John Bunker et al. [8] and the developments that flowed from it [30].

The absence of a treatment effect may be an important finding, but its importance may be missed if it is described as a "null result" with a P-value testing for difference above 0.05. Lacking evidence of an effect is not at all the same as possessing evidence that the effect is not large, yet a P-value above 0.05 in a test for difference is consistent with both of these situations. It is not possible to demonstrate a total absence of effect, but it is possible to demonstrate that the effect is not large using an equivalence test in a large randomized trial. In an observational study, there are added uncertainties because treatments were not randomly assigned. Nonetheless, in an observational study, one may find that the evidence that the effect is not large is insensitive to small or moderate biases from nonrandom treatment assignment.

[7] With two tests at $\Gamma = 1.4$, $\widetilde{H}_0$ is rejected but H_0 is not using the Bonferroni adjustment, but both $\widetilde{H}_0$ and H_0 are rejected using Holm's [18] procedure.

Table 23.1 Pain scores for surgical patients and controls in $I = 100$ matched pairs

Label	Group	Time	Minimum	Lower quartile	Median	Upper quartile	Maximum	Mean
a	Treated	Entry	0	2.0	5	6.0	9	4.28
b	Treated	Baseline	0	3.0	5	6.0	9	4.38
c	Treated	3 Months after surgery	0	1.8	3	5.0	9	3.55
d	Control	Entry	0	2.0	5	6.0	8	4.31
e	Control	Baseline	0	2.8	5	6.0	9	4.34
f	Control	3 Months later	0	2.0	3	5.0	8	3.39
Difference in changes			−7	−2.0	0	2.5	8	0.16

The change is pain at 3 months minus the average at entry and baseline, and the difference in changes refers to the treated-minus-control difference in these changes

To illustrate sensitivity analysis for an equivalence test, consider again the example in Sect. 13.2 and Table 13.1 concerning the effects of a surgical intervention, cystoscopy, and hydrodistention, on symptoms of interstitial cystitis (IC), a chronic urologic disorder characterized by bladder pain, and irritative voiding [26, 32]. In Table 13.1, the surgical treatment was estimated to have no effect on the nine-point pain score. How sensitive is this ostensible absence of effect to unmeasured biases from nonrandom treatment assignment?

Recall that pairs were formed using risk-set matching, matching at entry into the database and at baseline just prior to surgery for the surgical patients. Table 23.1 presents the pain scores in detail.[8] The surgically treated and control patients look similar in terms of pain at entry and prior to surgery for the surgical patient, and both groups look slightly and equally improved 3 months later. For the $I = 100$ pairs of two patients in Table 23.1, there are 400 "pretreatment" pain scores, merging "entry" and "baseline" scores; these have a mean of 4.3, a median of 5, a standard deviation of 2.2, and a median absolute deviation from the median of 2. Is it plausible that the effect of surgery is 2 units or more, but this is hidden from view in Table 13.1 because of bias in the assignment of patients to surgery or control?

The equivalence test applies Wilcoxon's signed rank statistic to test hypotheses about an additive treatment effect, $r_{Tij} - r_{Cij} = \tau$, with the null hypotheses asserting an effect that is not small, $H_{\neq}^{(\varsigma)} : |\tau| \geq \varsigma$, where $\varsigma > 0$ is set at 2 in this example; then, rejecting $H_{\neq}^{(\varsigma)}$ provides a basis for asserting with confidence that $|\tau| < \varsigma$. The hypothesis of inequivalence, $H_{\neq}^{(\varsigma)}$, is the union of two exclusive hypotheses, $\overleftarrow{H}_0^{(\varsigma)} : \tau \leq -\varsigma$ or $\overrightarrow{H}_0^{(\varsigma)} : \tau \geq \varsigma$. As a consequence, as in Sect. 23.3 and [2, 5, 6, 21, 36, 39, 45], $\overleftarrow{H}_0^{(\varsigma)}$ and $\overrightarrow{H}_0^{(\varsigma)}$ are each tested without a correction for multiple

[8] In [26], there is a sensitivity analysis for equivalence using all three outcomes at once in a coherent test. Here, only pain is considered.

Table 23.2 Sensitivity analysis for an equivalence test for the effect of surgery on pain scores in the IC data

Γ	$\overleftarrow{H}_0^{(2)}$	$\overrightarrow{H}_0^{(2)}$	$H_{\neq}^{(2)}$
1	4.9×10^{-9}	1.8×10^{-7}	1.8×10^{-7}
2	0.0012	0.0098	0.0098
2.4	0.0081	0.048	0.048
2.7	0.022	0.11	0.11
3	0.049	0.19	0.19

The table gives upper bounds on P-values. For an effect of 2 or more in either direction to be plausible, the magnitude of bias would need to exceed $\Gamma = 2.4$, and that magnitude of bias would only be able to mask an increase of 2 units in pain, not a decrease of 2 units in pain, caused by surgery. To mask a decrease of 2 units in pain caused by surgery, the bias would need to exceed $\Gamma = 3$

testing, and $H_{\neq}^{(\varsigma)}$ is rejected if both $\overleftarrow{H}_0^{(\varsigma)}$ and $\overrightarrow{H}_0^{(\varsigma)}$ are rejected. The sensitivity analysis for $H_{\neq}^{(\varsigma)}$ is then assembled from two standard sensitivity analyses, one for $\overleftarrow{H}_0^{(\varsigma)} : \tau \leq -\varsigma$, the other for $\overrightarrow{H}_0^{(\varsigma)} : \tau \geq \varsigma$; see [38] for a proof.

Table 23.2 is the sensitivity analysis for the equivalence test with $\varsigma = 2$. How much bias would need to be present for a moderately large treatment effect to appear as if it were no effect? More precisely: How much bias, Γ, would need to be present to produce the ostensible absence of effect in Table 13.1 if the treatment effect was 2 or more, $|\tau| \geq \varsigma = 2$? Table 23.2 gives upper bounds on P-values for the given values of Γ.[9] To mask an effect that is at least $|\tau| \geq 2$ units on the pain scale, the bias would need to be $\Gamma > 2.4$, and to mask a reduction in pain of at least $\tau \leq -2$, the bias would need to be $\Gamma > 3$. If the surgical procedure actually reduces pain by 2 units or more, the bias masking this effect would have to be fairly large.

[9]The upper bound on the P-value for $\overleftarrow{H}_0^{(\varsigma)} : \tau \leq -2$ is obtained by calculating Wilcoxon's signed rank statistic T from $Y_i + 2$ to test $\overrightarrow{H}_0^{(\varsigma)} : \tau = -2$ against $\tau > -2$ as in Sect. 3.5, rejecting if T is large. The upper bound on the P-value for $\overrightarrow{H}_0^{(2)} : \tau \geq 2$ is obtained by calculating Wilcoxon's signed rank statistic T from $Y_i - 2$ to test $\overrightarrow{H}_0^{(\varsigma)} : \tau = 2$ against $\tau < 2$ as in Sect. 3.5, rejecting if T is small. The P-value for testing $H_{\neq}^{(2)} : |\tau| \geq 2$ is the maximum of these two upper bounds; see [38] for a proof. The proof shows that the upper bound on the P-value for testing $H_{\neq}^{(2)} : |\tau| \geq 2$ is exactly equal to the maximum of the two separate P-values. As in Footnote 6 of this chapter, because $\overleftarrow{H}_0^{(\varsigma)}$ and $\overrightarrow{H}_0^{(\varsigma)}$ are exclusive, we can test both hypotheses at level α and yet take only an α risk of at least one false rejection. In consequence, we may reject either $\overleftarrow{H}_0^{(\varsigma)}$ or $\overrightarrow{H}_0^{(\varsigma)}$ even when we cannot reject both; see [36] and the appendix to this chapter.

23.6 Sensitivity Analysis for Equivalence and Difference

A planned analysis may seek to demonstrate a predicted positive difference for one outcome and predicted near equivalence for another, that is, an alternative hypothesis of superiority for one outcome and near equivalence for another [2, 42]. The demonstration would occur if a hypothesis of equality or inequivalence is rejected.[10]

A prediction of this sort arose in Sect. 13.3. In the study [38, 41] in Sects. 13.3 and 22.2 of early or late discharge of premature babies from the neonatal intensive care unit (NICU), there were 701 pairs of an early baby and a late baby who were similar on the day the early baby was discharged, although the late baby stayed in the hospital a few more days. Were the extra days in the NICU of benefit to the late babies?

The 6 months after the discharge of the early baby were divided into two intervals. During the first interval, the early baby was discharged home and the late baby was in the NICU. During the second interval, both babies had been discharged home. In [38, 41], health services provided to healthy babies (e.g., checkups) were set aside, and two outcomes in [38, 41] converted emergency and sick-baby health services into a dollar amount, one for the first interval, the other for the second interval. Because services were converted by schedule into costs, two babies who received the same services were coded as having the same costs; there is no variation in costs for the same service.

Let us devote a moment's thought to what patterns of outcomes might be seen and what they might mean. During the first interval, the late baby is in the NICU receiving hospital (or sick-baby) services, but the baby may be doing fine, growing older, chubbier, more mature, better able to face the world. During the first interval, the early baby is discharged home, so any emergency or sick-baby services are a sign that something has gone wrong, and if the dollar amount is very high, then something has gone very wrong, say a readmission through the emergency room. During the second interval, both babies are home, so sick-baby costs mean something has gone wrong. It is, therefore, important to distinguish the first and second intervals: for the late baby in the first interval, cost could be money and nothing more; elsewhere, cost is a sign that something has gone wrong. If substantial costs were at all common for the early babies in the first interval, this would suggest that the early discharges were premature. If early babies had higher second period costs, this too might suggest the early discharges were premature. If the late baby had consistently higher first period costs and second period costs were nearly equivalent, then the cost of retaining the late baby might plausibly have been

[10]This is a form of Roger Berger's [5] intersection–union test, in that one tests the disjunction of two hypotheses, $H_1 \vee H_2$, in an effort to demonstrate a conjunction, $(\sim H_1) \wedge (\sim H_2)$. Here, H_1 is equality or inferiority for one outcome and H_2 is inequivalence for another outcome, so superiority and equivalence is $(\sim H_1) \wedge (\sim H_2)$. In an intersection–union test, $H_1 \vee H_2$ is rejected at level α if H_1 and H_2 are rejected at level α; see [5, 25]. Intersection-union tests may be done in order, possibly rejecting H_1 when H_2 is not rejected, and a sensitivity analysis may be constructed from sensitivity analyses for the separate tests of H_1 and H_2; see the appendix to this chapter.

better spent in some other way, such as enhanced outpatient services for the babies. The analysis uses late-baby-minus-early-baby differences in costs for the first and second period.

The study concerned babies discharged alive. There were five deaths in the 6 months after discharge among these 1402 babies, or a rate of 3.6 per thousand babies per 6 months. In the USA as a whole in 2004, the first year mortality rate was 6.8 per thousand. These numbers are not comparable in several respects; one refers to 6 months, the other to a year; one refers to premature babies, the other to all babies; one refers to postdischarge mortality, the other includes deaths prior to discharge. For purposes of analysis, deaths were coded as infinite costs, that is, as the worst outcome [35]. There were three deaths (0.4%) among the early babies and two (0.3%) among the late babies. See [38, §2.1] for detailed discussion.

The Hodges–Lehmann point estimates (Sect. 2.4.3) of the typical difference in costs, late baby minus early baby, were \$4940 in the first period and \$17 in the second period, with 95% confidence intervals for an additive effect under the naïve model for treatment assignment (Sect. 13.5) of $\left[\$4485,\ \$5103\right]$ in the first period and $\left[-\$20,\ \$56\right]$ in the second period. Similar results were found using Stephenson's test (Sects. 2.5 and 17.1) which emphasizes consistent results in the extreme tails; see [38, Table 2]. All but 8/701 (1.1%) of early babies had less than \$300 of costs in the first period, and those 8 had costs in the range \$1422–\$9574. About 6% of late babies had first period costs above \$9556, but again the meaning of first period cost is different for early and late babies. The second period costs were typically low in both groups, with some extremely high costs, and no sign that late babies had lower second period costs; see [38, Table 1 and Figure 1].

Is the pattern of difference in the first period and equivalence in the second sensitive to unmeasured bias? The hypothesis is formulated in terms of additive effects τ_1 in period 1 and τ_2 in period 2. Inequivalence was defined as $|\tau_2| \geq \$500$, largely based on the thought that if you can be cured for less than \$500 in the USA, then you cannot be very sick. Recall that the Hodges–Lehmann estimate of the typical cost of a delayed discharge is about \$5000, and \$500 is 1/10 of that. The plan is to test for a difference in the first period, and if one is found, to test for equivalence in the second. Formally, the testing plan uses Wilcoxon's signed rank statistic to test $H_0 : \tau_1 \leq 0$, and if this hypothesis is rejected, the statistic is used to test $H_0 : |\tau_2| \geq \$500$. For each test separately, there is a sensitivity analysis, as in Sects. 3.5 and 23.5, and the upper bound on the P-value for testing $\tau_1 \leq 0$ or $|\tau_2| \geq \$500$—that is, for testing $H_0 : \tau_1 \leq 0 \vee |\tau_2| \geq \$500$—is the maximum of the upper bounds of the P-values for testing $H_0 : \tau_1 \leq 0$ and $H_0 : |\tau_2| \geq \$500$ separately, and, as in Sect. 23.5, the hypothesis $H_0 : |\tau_2| \geq \$500$ unpacks as either $H_0 : \tau_2 \leq -\$500$ or $H_0 : \tau_2 \geq \$500$.[11]

[11] This is true because the sequence of hypotheses $\langle H_1, \{H_2, H_3\}\rangle$ is a sequentially exclusive partition with H_1 asserting $\tau_1 \leq 0$, H_2 asserting $\tau_2 \leq -500$, and H_3 asserting $\tau_2 \geq 500$ because $\{H_2, H_2\}$ is exclusive. In fact, the three hypotheses may be tested in order, so H_1 may be rejected when H_2 and H_3 are not. See the appendix to the chapter for specifics.

Table 23.3 Sensitivity analysis for difference in first period costs (τ_1) and equivalence of second period costs (τ_2)

Γ	$H_0 : \tau_1 \leq 0$	$H_0 : \tau_2 \leq -\$500$	$H_0 : \tau_2 \geq \$500$	$H_0 : \tau_1 \leq 0 \vee \lvert\tau_2\rvert \geq \$500$
1	0.00001	0.00001	0.00001	0.00001
2	0.00001	0.00001	0.0020	0.0020
2.25	0.00001	0.00001	0.049	0.049
2.5	0.00001	0.00042	0.28	0.28
3	0.00001	0.056	0.90	0.90

Table 23.3 is the sensitivity analysis [38, Table 3]. Because $H_0 : \tau_1 \leq 0$ is rejected at the 0.05 level for each Γ in Table 23.3, the test plan continues and tests both $H_0 : \tau_2 \leq -\$500$ and $H_0 : \tau_2 \geq \$500$. For $\Gamma \leq 2.25$, all three hypotheses are rejected at the 0.05 level, so $H_0 : \tau_1 \leq 0 \vee |\tau_2| \geq \$500$ is rejected; that is, late discharge raised first period costs, but did not much affect second period costs. At $\Gamma = 2.5$, it is plausible that late discharge increased second period costs by \$500, but not plausible it reduced them by \$500. At $\Gamma = 3$, either an increase or a decrease of \$500 is plausible. It appears that late discharge raised first period costs, without compensating reduction in second period costs, and a small bias from nonrandom treatment assignment could not have created this appearance.

23.7 Summary

To disambiguate an association between treatment and response, Fisher advised "make your theories elaborate;" see Sect. 23.2 and [11, 33]. How should one plan the analysis of an observational study to evaluate the predictions of an elaborate theory? The plan should enlarge what can be said, not diminish it, yet control the frequency of false inferences. This is done by giving priority in the analysis to the evaluation of certain predictions, placing other predictions in a secondary role. The plan should distinguish predictions of a difference from predictions of near equivalence. This is implemented with appropriate use of equivalence tests and similar methods. The plan should permit a sensitivity analysis, because in an observational study it is never possible to be certain that adjustments for observed covariates have eliminated bias from covariates that were not measured.

23.8 Further Reading

Fisher's striking phrase "make your theories elaborate" has often been noted [11, 12, 14, 33].

There is a large literature about planning the analysis of clinical trials, and it provides some useful guidance for observational studies, where some of the issues

are the same and others are quite different. In particular, the confirmation, or partial confirmation, of an elaborate theory plays no role in a randomized experiment, where treatment assignment is known to be randomized.

The paper by Gary Koch and Stuart Gansky [22] is a gentle introduction to the clinical trials literature, whereas the review paper by Roger Berger and Jason Hsu [6] is a somewhat more technical discussion emphasizing tests of equivalence; see also the fine papers by Peter Bauer and colleagues [1–3]. The clinical trials literature quickly becomes technical and terse, so the serious reader will want to become acquainted with some of the early papers that shaped what came later, particularly the papers on "closed testing" by Ruth Marcus et al. [28], the intersection-union principle of Roger Berger [5], the device introduced by Julliet Popper Shaffer [40], and the false discovery rate of Yoav Benjamini and Yosef Hochberg [4]. Erich Lehmann's early paper [25] illuminates several issues and is of continuing interest. Two potentially important ideas that have received less attention are found in papers by Jason Hsu and Roger Berger [20] and Gerhard Hommel and Siegfried Kropf [19, §3]. Specifically, in [20], a multiple test procedure concludes with a confidence statement about a parameter, while in [19, §3] an ordered testing procedure continues past a few acceptances with the aid of a mild application of the Bonferroni inequality. See also [9, 46].

Jelle Goeman et al. [16] proposed a "three-sided test" that combines a two-sided test of the null hypothesis of no effect and two one-sided tests of null hypotheses of inequivalence. These three hypotheses are exclusive, so they may be tested without correction for multiple testing. Samuel Pimentel et al. [31] apply this method in an observational study of the surgical outcomes produced by surgeons who have just completed their formal surgical training. Do new surgeons have different outcomes than more experienced surgeons? Or are they nearly the same?

Less has been written about planning the analysis of an observational study. Issues in the current chapter are discussed in greater detail in [36, 38].

As discussed in this chapter, the Bonferroni inequality provides valid sensitivity bounds in observational studies, but they can be conservative. Individually sharp bounds on P-values may not be jointly sharp: perhaps no single bias in treatment assignment could simultaneously produce all of the upper bounds on the P-values. Colin Fogarty and Dylan Small [13] provide jointly sharp bounds on P-values, so use of Bonferroni inequality becomes less conservative. In their example, a bias of Γ could produce the observed association between one treatment and either of two outcomes; however, only a larger bias than Γ could produce both associations at the same time.

Appendix: Testing Hypotheses in Order

This chapter has considered several plans for testing one of Fisher's "elaborate theories" in a sequence of steps. These steps permit partial or total confirmation of the "elaborate theory." The investigator may select the first step in this sequence

on the basis of importance, priority, or conjecture. The first step in the sequence is an ordinary test done at level α as it might have been done without a plan for analysis, so in this rather specific sense, the subsequent steps cost the investigator nothing because the conclusion of the first step is the same. In brief, ambition in testing an "elaborate theory" is not penalized; at worst, the result is the partial confirmation that would have been obtained anyway in the single first test. In addition, these plans permitted a sensitivity analysis to be performed using the upper bounds on the P-values for the component tests; one applies the steps to the upper bounds. In other words, if you know how to do a sensitivity analysis for each step, then you know how to do a sensitivity analysis for the entire plan.

This appendix states and proves a result from [36], which then permits the statement of another result from [38]. The first result concerns conventional testing of a plan, while the second result concerns sensitivity analysis of the testing of a plan. Although the first result is tailored for use with an elaborate theory in an observational study, it has many analogues and precedents from the literature on multiple testing and closed testing [2, 19, 20, 22, 28, 40], equivalence testing [2, 6, 21, 39, 45], and multiparameter testing [5, 23, 25] in experiments. The second result takes more space to prove [38], so the proof is not reproduced here, but some discussion of it here is needed.

What Is a Sequentially Exclusive Partition of a Sequence of Hypotheses?

We have a collection of hypotheses, H_t for $t \in \mathscr{T}$, where $\mathscr{T}$ is a set of indices (or names) for the hypotheses. The set $\mathscr{T}$ is totally ordered (like a dictionary or telephone directory), and we write $t \prec t'$ if t is before t' in $\mathscr{T}$. To say that $\mathscr{T}$ is totally ordered is to say that it is always clear who is before whom: that is, if $t \in \mathscr{T}$ and $t' \in \mathscr{T}$ with $t \neq t'$ then either $t \prec t'$ or $t' \prec t$ (and never both $t \prec t'$ and $t' \prec t$). The order $\prec$ will partially determine which hypotheses are tested first, although other considerations will enter as well. Using numerical inequality, $<$, in place of $\prec$, five possible sets $\mathscr{T}$ are: $\mathscr{T} = \{1, 2, \ldots, k\}$, $\mathscr{T} = \{1, 2, \ldots\}$, $\mathscr{T} = \{\ldots, -1, 0, 1, 2, \ldots\}$, $\mathscr{T} = [0, \infty)$, and the positive rational numbers $\mathscr{T} = \{a/b : a, b \in \{1, 2, \ldots\}\}$. Other important totally ordered sets use the lexical order on a product of sets, as in the dictionary or telephone directory, where names are organized by first letter, then within first letter by second letter, and so on. The lexical order on $\mathscr{T} = \{1, 2, \ldots, k\} \times [0, \infty)$ says that infinitely many hypotheses $(1, a) \in \mathscr{T}$ with $a \in [0, \infty)$ come before $(2, 0)$. The lexical order on $\mathscr{T} = [0, \infty) \times \{1, 2, \ldots, k\}$ is very different: the k hypotheses $(0, a) \in \mathscr{T}$ come before all the rest. In Sect. 23.3, there were several orders that combined discrete and continuous aspects, testing one or two hypotheses, then building an interval estimate, or testing several hypotheses for infinitely many values of a parameter.

One assumption is made about the hypotheses, H_t for $t \in \mathscr{T}$, namely the "structure assumption." In practical terms, it is an innocuous assumption, but interestingly it seems to be needed nonetheless. A person disinclined to attend to innocuous assumptions might skip this paragraph. (While walking down a city street, a person disinclined to attend to innocuous assumptions might fall through a manhole whose cover is missing.) The structure assumption says: either all hypotheses H_t for $t \in \mathscr{T}$ are false or else there is a first true hypothesis. Stated precisely, either H_t is false for all $t \in \mathscr{T}$ or else there exists a $t' \in \mathscr{T}$ such that $H_{t'}$ is true and H_t is false for all $t \prec t'$. It takes a moment's thought to realize that the structure assumption is indeed an assumption: it might be false. For instance, let $\mathscr{T} = \{a/b : a, b \in \{1, 2, \ldots\}\}$ be the positive rational numbers ordered by inequality, $<$. Let H_t be the "hypothesis" that the circumference of a circle is less than or equal to t times its diameter. Then H_t is false for $t < \pi$ and is true for $t > \pi$, but there is not a first rational number $t > \pi$, so the structure assumption is false in this case. In this case, the structure assumption would be true if the positive rational numbers, $\mathscr{T} = \{a/b : a, b \in \{1, 2, \ldots\}\}$, were replaced by the positive real numbers, $\mathscr{T} = (0, \infty)$. The structure assumption seems to be a reminder that statistical inference needs the real numbers, and should avoid topological games involving things like the rational numbers; however, as far as I can see, the structure assumption is innocuous in practice because it is not an assumption about the world itself. Without further mention, the structure assumption is assumed to be true.

We are headed for a definition of a sequentially exclusive partition of hypotheses, a term that has been used informally at several points in this chapter. First we need to define an interval in $\mathscr{T}$ and then a partition of $\mathscr{T}$ into disjoint intervals.

It is natural to speak of an interval in $\mathscr{T}$. A nonempty subset $\mathscr{I} \subseteq \mathscr{T}$ is an interval if, for any $t,\ t',\ t'' \in \mathscr{T}$ with $t \prec t' \prec t''$, it follows that $t,\ t'' \in \mathscr{I}$ implies $t' \in \mathscr{I}$. For example, in $\mathscr{T} = \{1, 2, \ldots, 20\}$, the sets $\mathscr{I} = \{2\}$ and $\mathscr{I} = \{17, 18, 19\}$ are intervals, but $\{2, 4, 6\}$ is not an interval. In $\mathscr{T} = \{1, 2, \ldots, k\} \times [0, \infty)$ with the lexical order, the set $\{2\} \times [7, \infty) \bigcup \{3\} \times [0, 1]$ is an interval.

Disjoint intervals are ordered by their contents. If $\mathscr{I}_1 \subseteq \mathscr{T}$ and $\mathscr{I}_2 \subseteq \mathscr{T}$ are disjoint intervals, so $\mathscr{I}_1 \cap \mathscr{I}_2 = \emptyset$, then write $\mathscr{I}_1 \prec \mathscr{I}_2$ if $t \prec t'$ for all $t \in \mathscr{I}_1$ and $t' \in \mathscr{I}_2$. For instance, in $\mathscr{T} = \{1, 2, \ldots, 20\}$, we have $\{2, 3\} \prec \{6, 7, 8\}$. Write $t \prec \mathscr{I}$ if $\{t\} \prec \mathscr{I}$ and $\mathscr{I} \prec t$ if $\mathscr{I} \prec \{t\}$. For instance, in $\mathscr{T} = \{1, 2, \ldots, 20\}$, we have $2 \prec \{6, 7, 8\}$.

A partition of $\mathscr{T}$ into disjoint intervals is a representation of $\mathscr{T}$ as the union of disjoint intervals $\mathscr{I}_\lambda$, indexed by a set Λ, that is, $\mathscr{T} = \bigcup_{\lambda \in \Lambda} \mathscr{I}_\lambda$ with $\mathscr{I}_\lambda \cap \mathscr{I}_{\lambda'} = \emptyset$ for $\lambda, \lambda' \in \Lambda$ with $\lambda \neq \lambda'$. In $\mathscr{T} = \{1, 2, \ldots, 20\}$, with $\Lambda = \{a, b, c\}$, one partition of $\mathscr{T}$ into disjoint intervals is $\mathscr{T} = \bigcup_{\lambda \in \Lambda} \mathscr{I}_\lambda$ where $\mathscr{I}_a = \{1, 2, \ldots, 9\}$, $\mathscr{I}_b = \{10\}$, $\mathscr{I}_c = \{11, 12, \ldots, 20\}$. Because the intervals, $\mathscr{I}_\lambda$, $\lambda \in \Lambda$, are disjoint, they are ordered; so write $\lambda \prec \lambda'$ if $\mathscr{I}_\lambda \prec \mathscr{I}_{\lambda'}$, and for $t \in \mathscr{T}$, write $t \prec \lambda$ if $t \prec \mathscr{I}_\lambda$ or $\lambda \prec t$ if $\mathscr{I}_\lambda \prec t$. For instance, in the partition of $\mathscr{T} = \{1, 2, \ldots, 20\}$ just mentioned, $\mathscr{I}_a \prec \mathscr{I}_c$ and $7 \prec \mathscr{I}_c$.

A set $\mathscr{S} \subseteq \mathscr{T}$ of hypotheses is exclusive if at most one of the hypotheses H_t, $t \in \mathscr{S}$, is true. For instance, if θ is a real parameter, then the set of hypotheses $\{H_t : t \in (-\infty, \infty)\}$ would be exclusive if H_t were $H_t : \theta = t$, but the same set would not be exclusive if the H_t were $H_t : \theta \leq t$; in the first case, distinct

hypotheses contradict each other but in the second case they do not. The words "at most one" are important: the set of hypotheses $\{H_t : t \in (12,\ 17)\}$ with $H_t : \theta = t$ is exclusive even if there is no reason to think θ is between 12 and 17.

Now, we can define a sequentially exclusive partition of hypotheses. A partition of $\mathscr{T}$ into disjoint intervals, $\mathscr{T} = \bigcup_{\lambda \in \Lambda} \mathscr{I}_\lambda$, is a sequentially exclusive partition if each interval $\mathscr{I}_\lambda$, $\lambda \in \Lambda$, would be exclusive if all hypotheses in earlier intervals were false; that is, at most one hypothesis H_t, $t \in \mathscr{I}_\lambda$, is true whenever all hypotheses, $H_{t'}$, $t' \prec \mathscr{I}_\lambda$, are false. For instance, with $\mathscr{T} = \{1, 2, 3\}$, the partition $\mathscr{T} = \{1\} \cup \{2, 3\}$ is a sequentially exclusive partition if at most one of H_2 and H_3 is true whenever H_1 is false; see Sect. 23.3 for several examples. Notice that $\{2, 3\}$ need not be exclusive if H_1 is true.

A sequentially exclusive partition always exists, namely the partition $\mathscr{T} = \bigcup_{t \in \mathscr{T}} \{t\}$. This is so because a set containing a single hypothesis is always exclusive; it contains at most one true hypothesis.

Testing Hypotheses in Order

In the following testing plan, $\mathscr{T} = \bigcup_{\lambda \in \Lambda} \mathscr{I}_\lambda$ is a sequentially exclusive partition, and for each hypothesis, H_t, $t \in \mathscr{T}$, there is a valid P-value, p_t, such that $\Pr(p_t \leq \alpha) \leq \alpha$ for all $\alpha \in [0, 1]$ if H_t is true. Fix an α, conventionally $\alpha = 0.05$.

Testing plan. If $p_s \leq \alpha$ for all $s \in \mathscr{I}_\lambda$ for all $\lambda \prec \omega$, then reject all hypotheses H_t, $t \in \mathscr{I}_\omega$, that have $p_t \leq \alpha$.

This testing plan is, in fact, the general form of the plan used throughout this chapter. For the partition $\mathscr{T} = \bigcup_{t \in \mathscr{T}} \{t\}$, the testing plan says to reject all hypotheses prior to the first acceptance and stop. In general, with $\mathscr{T} = \bigcup_{\lambda \in \Lambda} \mathscr{I}_\lambda$, the procedure stops with the first interval $\mathscr{I}_\omega$ that contains at least one hypothesis $t' \in \mathscr{I}_\omega$ with $p_{t'} > \alpha$; however, all hypotheses $t \in \mathscr{I}_\omega$ are tested, and all of those with $p_t \leq \alpha$ are rejected. Particular cases of Proposition 23.1 have been discussed throughout this chapter.

Proposition 23.1 ([36]) *If the testing plan is applied to a sequentially exclusive partition of hypotheses, H_t, $t \in \mathscr{T} = \bigcup_{\lambda \in \Lambda} \mathscr{I}_\lambda$, where $\Pr(p_t \leq \alpha) \leq \alpha$ if H_t is true, then the probability that the plan tests and rejects at least one true hypothesis is at most α.*

Proof Let $\mathscr{W}$ be the event that the testing plan tests and rejects at least one true hypothesis. If all hypotheses, H_t, $t \in \mathscr{T}$, are false, then there is nothing to prove; in particular, $\Pr(\mathscr{W}) = 0$. Otherwise, let H_v be the first true hypothesis, let $\mathscr{R}_v$ be the event that the testing plan tests and rejects H_v, and let $\mathscr{I}_\omega$ be the unique interval that contains v. Then H_t is false for all $t \prec \mathscr{I}_\omega$. This implies that H_v is the only true hypothesis in $\mathscr{I}_\omega$, because $\mathscr{T} = \bigcup_{\lambda \in \Lambda} \mathscr{I}_\lambda$ is sequentially exclusive. It follows that the plan tests and rejects at least one true hypothesis if and only if it tests and rejects H_v, so $\mathscr{W} = \mathscr{R}_v$, and $\Pr(\mathscr{W}) = \Pr(\mathscr{R}_v) \leq \alpha$.

Sensitivity Analysis for Testing in Order

In Sect. 3.5, a sensitivity analysis was performed in which an upper bound, say $P_{t,\max}$, was found for the one-sided P-value when testing a single hypothesis, say H_t; see Table 3.2. This upper bound, $P_{t,\max}$, was the largest P-value that could be produced for a given magnitude Γ of deviation from random treatment assignment in (3.13). That bound was sharp, in the sense that there was some set of treatment assignment probabilities π_{ij} satisfying (3.16)–(3.18) for which the bound $P_{t,\max}$ was the correct P-value.

Suppose that we have a sequence of such bounds, $P_{t,\max}$, for a sequence of hypotheses, H_t, $t \in \mathscr{T}$, where $\mathscr{T} = \bigcup_{\lambda \in \Lambda} \mathscr{I}_\lambda$ is a sequentially exclusive partition of hypotheses. Suppose that the testing plan was applied using the bounds in place of the unknown, actual P-values. What would happen? In fact, this was done several times in this chapter.

It is possible to show the following [38, Proposition 1]. First, if (3.13) is true for a given value of $\Gamma \geq 1$, the chance that the testing plan tests and rejects at least one true hypothesis is at most α. Second, this bound is sharp: there are treatment assignment probabilities π_{ij} satisfying (3.16)–(3.18) for which the testing plan would terminate with acceptance of the same hypothesis, H_v, in the proof of Proposition 23.1. In other words, if the testing plan was used once for every set of treatment assignment probabilities π_{ij} satisfying (3.16)–(3.18), then a hypothesis H_t would be rejected in all of these analyses if and only if it is rejected in the one analysis that uses the individual bounds in place of the actual P-values.

The sharpness of the bound is actually somewhat surprising. The reason it is surprising is that the treatment assignment probabilities π_{ij} that produce $P_{t,\max}$ are not generally the same as the π_{ij}'s that produce $P_{t',\max}$ for $t \neq t'$, so there may be no treatment assignment probabilities satisfying (3.16)–(3.18) that produce both $P_{t,\max}$ and $P_{t',\max}$. You can have $P_{t,\max}$ or you can have $P_{t',\max}$, but there is no assurance that you can have both $P_{t,\max}$ and $P_{t',\max}$, even though the proposal is to plug them both into the testing plan as if they could coexist. If the Bonferroni inequality is used with $P_{t,\max}$ and $P_{t',\max}$, then the result is conservative, not sharp [38, §4.5]: it can happen that $\min\left(P_{t,\max},\ P_{t',\max}\right) > \alpha/2$, while the minimum of the two P-values for testing H_t and $H_{t'}$ is less than $\alpha/2$ for all treatment assignment probabilities π_{ij} satisfying (3.16)–(3.18). Speaking informally, the bound is sharp for the testing plan because the plan stops with a single P-value; see [38, Proposition 1] for a precise statement.

The results in [36, 38] are general; they are not restricted to matched pairs.

Bibliography

1. Bauer, P.: Multiple testing in clinical trials. Stat. Med. **10**, 871–890 (1991)
2. Bauer, P., Kieser, M.: A unifying approach for confidence intervals and testing of equivalence and difference. Biometrika **83**, 934–937 (1996)

3. Bauer, P.: A note on multiple testing procedures in dose finding. Biometrics **53**, 1125–1128 (1997)
4. Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate. J. R. Stat. Soc. B **57**, 289–300 (1995)
5. Berger, R.L.: Multiparameter hypothesis testing and acceptance sampling. Technometrics **24**, 295–300 (1982)
6. Berger, R.L., Hsu, J.C.: Bioequivalence trials, intersection-union tests and equivalence confidence sets. Stat. Sci. **11**, 283–319 (1996)
7. Bilban, M., Jakopin, C.B.: Incidence of cytogenetic damage in lead-zinc mine workers exposed to radon. Mutagenesis **20**, 187–191 (2005)
8. Bunker, J.P., Barnes, B.A., Mosteller, F.: Costs, Risks and Benefits of Surgery. Oxford University Press, Oxford (1977)
9. Burman, C.F., Sonesson, C., Guilbaud, O.: A recycling framework for the construction of Bonferroni-based multiple tests. Stat. Med. **28**, 739–761 (2009)
10. Caughey, D., Dafoe, A., Seawright, J.: Nonparametric combination (NPC): a framework for testing elaborate theories. J. Politics **79**, 688–701 (2017)
11. Cochran, W.G.: The planning of observational studies of human populations (with Discussion). J. R. Stat. Soc. A **128**, 234–265 (1965)
12. Cox, D.R.: Causality: some statistical aspects. J. R. Stat. Soc. A **155**, 291–301 (1992)
13. Fogarty, C.B., Small, D.S.: Sensitivity analysis for multiple comparisons in matched observational studies through quadratically constrained linear programming. J. Am. Stat. Assoc. **111**, 1820–1830 (2016)
14. Gail, M.: Statistics in action. J. Am. Stat. Assoc. **91**, 1–13 (1996)
15. Goeman, J.J., Solari, A.: The sequential rejection principle of familywise error control. Ann. Stat. **38**, 3782–3810 (2010)
16. Goeman, J.J., Solari, A. Stijnen, T.: Three-sided hypothesis testing: simultaneous testing of superiority, equivalence and inferiority. Stat. Med. **29**, 2117–2125 (2010)
17. Heller, R., Rosenbaum, P.R., Small, D.: Split samples and design sensitivity in observational studies. J. Am. Stat. Assoc. **104**, 1090–1101 (2009)
18. Holm, S.: A simple sequentially rejective multiple test procedure. Scand. J. Stat. **6**, 65–70 (1979)
19. Hommel, G., Kropf, S.: Tests for differentiation in gene expression using a data-driven order or weights for hypotheses. Biomet. J. **47**, 554–562 (2005)
20. Hsu, J.C., Berger, R.L.: Stepwise confidence intervals without multiplicity adjustment for dose-response and toxicity studies. J. Am. Stat. Assoc. **94**, 468–475 (1999)
21. Hsu, J.C., Hwang, J.T.G., Liu, H-K., Ruberg, S.J.: Confidence intervals associated with tests for bioequivalence. Biometrika **81**, 103–114 (1994)
22. Koch, G.G., Gansky, S.A.: Statistical considerations for multiplicity in confirmatory protocols. Drug. Inform. J. **30**, 523–533 (1996)
23. Laska, E.M., Meisner, M.J.: Testing whether an identified treatment is best. Biometrics **45**, 1139–1151 (1989)
24. Lehmacher, W., Wassmer, G., Reitmeir, P.: Procedures for two-sample comparisons with multiple endpoints controlling the experimentwise error rate. Biometrics **47**, 511–521 (1991)
25. Lehmann, E.L.: Testing multiparameter hypotheses. Ann. Math. Stat. **23**, 541–552 (1952)
26. Li, Y.F.P., Propert, K.J., Rosenbaum, P.R.: Balanced risk set matching. J. Am. Stat. Assoc. **96**, 870–882 (2001)
27. Lu, X., White, H.: Robustness checks and robustness tests in applied economics. J. Econ. **178**, 194–206 (2014)
28. Marcus, R., Peritz, E., Gabriel, K.R.: On closed testing procedures with special reference to ordered analysis of variance. Biometrika **63**, 655–60 (1976)
29. Masjedi, M.R., Heidary, A., Mohammadi, F., Velayati, A.A., Dokouhaki, P.: Chromosomal aberrations and micronuclei in lymphocytes of patients before and after exposure to anti-tuberculosis drugs. Mutagenesis **15**, 489–494 (2000)

30. McPherson, K., Bunker, J.P.: Costs, risks and benefits of surgery: a milestone in the development of health services research. J. R. Soc. Med. **100**, 387–390 (2007)
31. Pimentel, S.D., Kelz, R.R., Silber, J.H., Rosenbaum, P.R.: Large, sparse optimal matching with refined covariate balance in an observational study of the health outcomes produced by new surgeons. J. Am. Stat. Assoc. **110**, 515–527 (2015)
32. Propert, K.J., Schaeffer, A.J., Brensinger, C.M., Kusek, J.W., Nyberg, L.M., Landis, J.R.: A prospective study of interstitial cystitis: results of longitudinal followup of the interstitial cystitis data base cohort. J. Urol. **163**, 1434–1439 (2000)
33. Rosenbaum, P.R.: From association to causation in observational studies. J. Am. Stat. Assoc. **79**, 41–48 (1984)
34. Rosenbaum, P.R.: Observational Studies, 2nd edn. Springer, New York (2002)
35. Rosenbaum, P.R.: Comment on a paper by Donald B. Rubin: the place of death in the quality of life. Stat. Sci. **21**, 313–316 (2006)
36. Rosenbaum, P.R.: Testing hypotheses in order. Biometrika **95**, 248–252 (2008)
37. Rosenbaum, P.R.: Combining planned and discovered comparisons in observational studies. Biostatistics **21**, 384–399 (2020). https://doi.org/10.1093/biostatistics/kxy055
38. Rosenbaum, P.R., Silber, J.H.: Sensitivity analysis for equivalence and difference in an observational study of neonatal intensive care units. J. Am. Stat. Assoc. **104**, 501–511 (2009)
39. Schuirmann, D.L.: On hypothesis testing to determine if the mean of a normal distribution is contained in a known interval. Biometrics **37**, 617 (1981)
40. Shaffer, J.P.: Modified sequentially rejective multiple test procedures. J. Am. Stat. Assoc. **81**, 826–831 (1986)
41. Silber, J.H., Lorch, S.L., Rosenbaum, P.R., Medoff-Cooper, B., Bakewell-Sachs, S., Millman, A., Mi, L., Even-Shoshan, O., Escobar, G.E.: Additional maturity at discharge and subsequent health care costs. Health Serv. Res. **44**, 444–463 (2009)
42. Tamhane, A., Logan, B.: A superiority-equivalence approach to one-sided tests on multiple endpoints in clinical trials. Biometrika **91**, 715–727 (2004)
43. Tukey, J.W.: We need both exploratory and confirmatory. Am. Stat. **34**, 23–25 (1980)
44. Tukey, J.W.: Sunset salvo. Am. Stat. **40**, 72–76 (1986)
45. Westlake, W.J.: Response to Kirkwood. Biometrics **37**, 591–593 (1981)
46. Wiens, B.L., Dmitrienko, A.: The fallback procedure for evaluating a single family of hypotheses. J. Biopharm. Stat. **15**, 929–942 (2005)
47. Yoon, F.: New methods for the design and analysis of observational studies. Doctoral Thesis, Department of Statistics, University of Pennsylvania (2009)
48. Yoon, F.B., Huskamp, H.A., Busch, A.B., Normand, S.L.T.: Using multiple control groups and matching to address unobserved biases in comparative effectiveness research: an observational study of the effectiveness of mental health parity. Stat. Biosci. **3**, 63–78 (2011)
49. Zhao, Q., Small, D.S., Rosenbaum, P.R.: Cross-screening in observational studies that test many hypotheses. J. Am. Stat. Assoc. **113**, 1070–1084 (2018)

Summary: Key Elements of Design

In an Observational Study, Competing Theories Should Make Conflicting Predictions Many studies dissipate before they begin from a simple lack of focus. The study is intended to settle something, or at least take a step towards settling something, and for that, there needs to be something definite to settle and some prospect of and means for settling it; see Chap. 4. Some pairs of theories rarely yield conflicting predictions, so one may search for unusual opportunities to contrast them; see Sect. 5.1.

An Observational Study Should Be Structured to Resemble a Simple Experiment A typical structure is the comparison of a treated and a control group that looked comparable prior to treatment in terms of observed covariates. An experiment is an unusual situation. In a carefully controlled experiment, one of the rarest of things happens: the effects caused by treatments are seen with clarity. Each step away from the experimental template is a step closer to the edge of an abyss; see Sects. 1.2 and 13.1.

Adjustments for Observed Covariates Should Be Simple, Transparent, and Convincing The major source of uncertainty about the conclusions of an observational study comes from the possible failure to control for covariates that were not measured. This possibility is raised in virtually every observational study. If you think this possibility will not be raised in evaluating your study, then you are kidding yourself. There is little hope of addressing this major source of uncertainty if the study becomes bogged down in unnecessarily complex, obscure, or unconvincing adjustments for observed covariates. One simple, transparent, and convincing way to adjust for observed covariates is to compare a treated and a control group with similar distributions of the observed covariates. Such a control group may often be constructed with the aid of multivariate matching; see Part II.

The Most Plausible Alternatives to a Treatment Effect Should Be Anticipated and Addressed In an observational study, it is not possible to address every

© Springer Nature Switzerland AG 2020

P. R. Rosenbaum, *Design of Observational Studies*, Springer Series in Statistics,
https://doi.org/10.1007/978-3-030-46405-9

conceivable alternative to a treatment effect. It is often possible to anticipate several plausible objections or counterclaims to a claim that a comparison of matched treated and control groups estimates the effects of the treatment. Such an objection claims that the comparison is ambiguous, that it could estimate a treatment effect or it could be distorted by some specific form of bias. With a specific form of bias in mind, design elements can often be added, such as two control groups or unaffected outcomes, that resolve specific ambiguities; see Sect. 5.2.

The Analysis Should Address Possible Biases from Unmeasured Covariates Typically, the analysis should include a sensitivity analysis of one form or another; see Chap. 3. A sensitivity analysis asks: how much bias from unmeasured covariates—what magnitude of deviation from random assignment—would need to be present to qualitatively alter the conclusions suggested by the naïve, straightforward comparison of matched treated and control groups? The degree of sensitivity to unmeasured bias is a fact of the matter, something that is determined without ambiguity from the data at hand. Whether or not biases of this magnitude are present remains a matter of reasoned conjecture and responsible debate, but that debate is now informed and constrained by the facts of the matter. A P-value does not rule out the possibility that bad luck produced the observed results; rather, it objectively measures how much bad luck would be required to produce the observed results. A sensitivity analysis does not rule out the possibility that unmeasured bias produced the observed results; rather, it objectively measures how much unmeasured bias would be required to produce the observed results.

To the Extent Possible, Observational Studies Should Be Designed to Be Insensitive to Biases from Unmeasured Covariates To do this, one must know what makes some designs (or data generating processes) sensitive to unmeasured biases and others insensitive. With this knowledge, when faced with a choice, an insensitive design may be chosen. Many factors strongly affect design sensitivity; see Part III.

There Should Be a Plan for a Primary Analysis A randomized controlled trial invariably has a protocol detailing a plan for a primary analysis. An observational study also needs such a plan. With design elements intended to resolve otherwise ambiguous comparisons, the study design anticipates one of a few patterns of results, what R.A. Fisher called an "elaborate theory"; see Sect. 23.2. The typical elaborate theory predicts a difference here, near equivalence there, the absence of a trend here, a discontinuity there. For instance, an elaborate theory might predict much higher responses in the treated group than in two control groups, with the two control groups differing negligibly from each other. A planned primary analysis will attempt to examine, possibly confirm, the predictions of an elaborate theory; see Chap. 23. The predictions of an elaborate theory are predictions only if they precede examination of the outcomes; there is no value in an elaborate theory constructed after the fact to accommodate a particular set of data. Therefore, the analysis of an elaborate theory must be a planned analysis. Planning may be aided by sample splitting; see Sect. 22.1. A plan for a primary analysis does not preclude unplanned exploratory analyses; rather, it distinguishes planned and unplanned analyses.

Solutions to Common Problems

Solutions are offered to common problems that arise in the design of observational studies. The discussion here is a problem-oriented guide to the book as a whole. Starting with a common problem, a brief solution is described; then, relevant sections of the book are identified. Also discussed are certain situations that are often perceived as problematic, but are not actually problems.

The Matching Problem Is Too Large If a matching problem is too large, divide it into several smaller problems by exact matching for one or more important covariates; see Sect. 10.4. Alternatively, consider the methods in [12].

Treated and Control Groups Are Too Far Apart to Match Before trying to solve this problem, make sure the problem is real. Compare boxplots of the propensity scores for treated and control groups. If the boxplots cover much the same range, but some covariates or the propensity score are poorly balanced, consider: (1) tightening the caliper on the propensity score so that, for example, the penalty engages at 10% of the standard deviation rather than 20% (see Sect. 9.4), (2) use penalties to improve balance on one or two stubborn covariates (see Sect. 10.3), or (3) try full matching (see Sect. 9.6). Otherwise, if the boxplots exhibit large regions with little or no overlap, consider redefining the study population using a few key covariates (see Sect. 3.7). For examples of redefining the study population, see [1, 2].

Treated and Control Groups Overlap, but at Some Values of x There Are Too Few Controls Even for 1-to-1 Matching Try full matching; see Sect. 9.6.

People Are Treated at Different Times. How Do I Match? Consider risk-set matching; see Chap. 13.

I Want to Match for a Variable with Many Categories, but There Are Not Enough People in the Categories to Permit a Close Match Try matching with fine balance; see Chap. 11.

© Springer Nature Switzerland AG 2020

P. R. Rosenbaum, *Design of Observational Studies*, Springer Series in Statistics,
https://doi.org/10.1007/978-3-030-46405-9

I Have Two Control Groups. How Do I Match? There are several options. One approach is to form matched triples; see Chap. 21. Another approach forms matched pairs in an "incomplete block" design; see Chap. 12. A different problem is to split one control group to form two for a specific purpose; see the discussion of tapered matching and the exterior match in Chap. 21.

My Propensity Score Model Does a Poor Job of Predicting Treatment Assignment Not a problem. In a large, completely randomized experiment, a propensity score model would have great difficulty predicting treatment assignments from covariates precisely because treatment assignment is random and does not depend upon the covariates. The propensity score is intended to fix a specific problem, namely imbalances in observed covariates. If your study does not have that problem, then that is just fine.

How Do I Judge Whether My Model for the Propensity Score Is a Good Model? The propensity score has various uses, and the answer to this question depends upon how the propensity score will be used. In this book, propensity scores are used for matching. When used for matching, propensity scores are a means to an end, namely matched pairs or sets that balance observed covariates. When you have matched pairs or sets that balance observed covariates, the matching for observed covariates is done, and attention shifts to potential bias from unmeasured covariates. In light of this, judging the propensity score model when used in matching is essentially the same task as judging whether the matching has balanced observed covariates; see Sect. 10.1.

How Do I Select Covariates to Use in the Propensity Score? This question inverts the means and the end. The proper question is: Which covariates do you wish to balance by matching on the propensity score?

There Is a Covariate that Strongly Predicts Treatment Assignment Z but Seems Unlikely to Matter Much for the Response R. What Should I Do? Read about "seemingly innocuous confounding" and analytical adjustments in Sect. 22.2 and about tapered matching and the exterior match in Chap. 21.

There Is a Variable that Is Subsequent to Treatment Assignment and May Have Been Affected by the Treatment, so It Is Not a Covariate, but I Feel I Should Adjust for It Anyway Although this is sometimes reasonable, think long and hard before you do this. If you adjust for a concomitant variable that has been affected by the treatment, you may introduce a bias that would not otherwise have been present; see [5]. If in doubt, it may be best to leave such a variable unmatched, so that analyses with and without analytical adjustments for it are possible; see Sect. 22.2. One alternative is to both match and not match for such a variable; see the discussion of tapered matching and the exterior match in Chap. 21.

I Have Matched with a Variable Ratio of Controls to Treated Subjects, and Now I Want a Boxplot It is not difficult to do, but at the time I write this, current software does not do it, so a few steps are needed. Essentially, you need to compute a weighted empirical distribution function, compute the quantiles from it, and make a

boxplot using those quantiles. The R function "bxp" will help: it will draw boxplots from quantiles you give it. See Note 3 in Chap. 9. The resulting boxplots look like ordinary boxplots, but they are weighted to reflect the variable numbers of controls. Examples are found in [2].

Should I Match with a Fixed or a Variable Ratio of Controls to Treated Subjects? The choice is discussed in Sect. 9.5. The previous solution, about the boxplot, illustrates the biggest disadvantage of matching with variable controls; simple, straightforward tasks require special programming effort. Opposed to this, theory strongly suggests that matching with variable controls is an efficient way to produce closer matches; see [4]. Matching with an average of two controls per treated individual is often feasible when matching every treated individual to exactly two controls is not. Multiple controls can increase the design sensitivity [9].

People Say that a Specific Unobserved Covariate u Is Strongly Related to the Outcome, r_C, but I Doubt It To show that a specific unobserved covariate u is not strongly related to the outcome r_C, find two control groups that differ markedly in terms of the unobserved u, and show that outcomes do not greatly differ in these two control groups. See Sects. 5.2.2, 23.3 and Chap. 21.

People Say that a Specific Unobserved Covariate u Is Strongly Related to Treatment Assignment Z, but I Doubt It To show that a specific unobserved covariate u is not strongly related to treatment assignment, Z, find an outcome known to be unaffected by the treatment that is highly correlated with u, and show that this unaffected outcome has a similar distribution in treated ($Z = 1$) and control ($Z = 0$) groups. See Sect. 5.2.4.

I Would Like to Perform a Sensitivity Analysis, but I Do Not Have Matched Pairs It is not difficult. See Sect. 19.7, [6, Chapter 4] and [11].

I Would Like to Perform a Sensitivity Analysis, but My Outcomes Are Binary Again, it is not difficult. See [6, Chapter 4]. That chapter also discusses outcomes that are censored survival times.

How Do I Interpret the Parameter Γ? The parameter Γ is convenient in that it is a single parameter that can refer to a wide variety of situations. The resulting sensitivity analysis is one-dimensional: just one parameter varies. The sensitivity bounds implicitly refer to an unobserved covariate u that is strongly related to the response, r_C, and that has a controlled relationship with the treatment, Z, the control being provided by the value of Γ. Sometimes, that situation is not the one under discussion, because a very strong relationship between u and r_C is not plausible. It is possible to reexpress the one-parameter analysis in terms of two parameters, where one parameter controls the relationship between the unobserved covariate u and the response r_C and the other controls the relationship between u and the treatment, Z. It is the same one-dimensional sensitivity analysis, with no new computations, but now with a two-dimensional interpretation; see Sect. 3.6, [7] and [10, Table 9.1].

I Am Worried that My Findings Will Be Sensitive to Small Unmeasured Biases Review Chaps. 15–19 to see what issues affect the sensitivity of conclusions to unmeasured biases. Consider use of evidence factors to increase insensitivity to unmeasured biases; see Chap. 20. Consider sample splitting to guide design decisions that affect design sensitivity; see Sect. 22.1 and [3].

I Have Doses of Treatment, but I Am Not Sure Whether They Are Any Good Split the sample and find out; see Sect. 22.1 and [3]. Alternatively, use adaptive inference to gain the benefits of the better analyses; see Sect. 19.3 and [8]. Similar advice applies to many other design decisions.

The Treatment that Interests Me Is Known to Be Assigned in a Very Biased Fashion Instead of studying the effect of that treatment, consider studying its differential effect compared with other treatments affected by the same biases; see Sect. 5.2.6. This may help in some circumstances.

There Have Been Several Studies of the Treatment that Interests Me, but None Is Convincing Is it possible to study the treatment again, this time removing one of the several problems that made previous studies unconvincing? See Sect. 4.5.

The Same Problems Inevitably Occur Whenever Efforts Are Made to Study the Treatment that Interests Me Why is the treatment thought to work? What reasons are given? Why is there doubt about the effects of the treatment? What reasons are given? Can an empirical study shed light on whether these reasons are valid? Perhaps the reasons can be supported or refuted, even though direct investigation of the effects is difficult. See Sect. 4.6.

I Am Disappointed by My "Null Result" Do you possess evidence for the absence of a large effect? That might be an important finding; see Sect. 23.5. Or do you lack evidence about the magnitude of the effect? That is disappointing, but it is a common occurrence, one that happens now and then to everyone.

References

1. Fogarty, C.B., Mikkelsen, M.E., Gaieski, D.F., Small, D.S.: Discrete optimization for interpretable study populations and randomization inference in an observational study of severe sepsis mortality. J. Am. Stat. Assoc. **111**, 447–458 (2016)
2. Haviland, A.M., Nagin, D.S., Rosenbaum, P.R.: Combining propensity score matching and group-based trajectory analysis in an observational study. Psychol. Methods **12**, 247–267 (2007)
3. Heller, R., Rosenbaum, P.R., Small, D.: Split samples and design sensitivity in observational studies. J. Am. Stat. Assoc. **104**, 1090–1101 (2009)
4. Ming, K., Rosenbaum, P.R.: Substantial gains in bias reduction from matching with a variable number of controls. Biometrics **56**, 118–124 (2000)
5. Rosenbaum, P.R.: The consequences of adjustment for a concomitant variable that has been affected by the treatment. J. R. Stat. Soc. A **147**, 656–666 (1984)
6. Rosenbaum, P.R.: Observational Studies, 2nd edn. Springer, New York (2002)

7. Rosenbaum, P.R., Silber, J.H.: Amplification of sensitivity analysis in observational studies. J. Am. Stat. Assoc. **104**, 1398–1405 (2009)
8. Rosenbaum, P.R.: Testing one hypothesis twice in observational studies. Biometrika **99**, 763–774 (2012)
9. Rosenbaum, P.R.: Impact of multiple matched controls on design sensitivity in observational studies. Biometrics **69**, 118–127 (2013)
10. Rosenbaum, P.R.: Observation and Experiment: An Introduction to Causal Inference. Harvard University Press, Cambridge (2017)
11. Rosenbaum, P.R.: Sensitivity analysis for stratified comparisons in an observational study of the effect of smoking on homocysteine levels. Ann. Appl. Stat. **12**, 2312–2334 (2018)
12. Yu, R., Silber, J.H., Rosenbaum, P.R.: Matching methods for observational studies derived from large administrative databases. Stat. Sci. (2020, in press)

Symbols

$\lvert A \rvert$	If A is a finite set, $\lvert A \rvert$ is the number of elements in A
$\mathbf{x}$	Observed covariates for any one person. See Sect. 2.1.2
$\mathbf{x}_\ell$	Observed covariates for person ℓ, $\ell = 1, \ldots, L$. The ℓ subscript is used to denote people before they are matched. The ℓ subscript is used in the same way with Z, R, r_T, r_C, and u. See Sect. 3.1
$\mathbf{x}_{ij}$	Observed covariates for person j, in pair i, $i = 1, 2, \ldots, I$, $j = 1, 2$. The ij subscripts are used to denote people after they are arranged into matched pairs. The ij subscripts are used in the same way with Z, R, r_T, r_C, and u. See Sect. 2.1.2
Z	Treatment indicator, $Z = 1$ if treated, $Z = 0$ if control. See Sect. 2.1.2
$\mathbf{Z}$	Treatment indicators for the $2I$ subjects in the I matched pairs, $\mathbf{Z} = (Z_{11}, Z_{12}, Z_{21}, \ldots, Z_{I2})^T$. See Sect. 2.1.2
$e(\mathbf{x})$	Propensity score, $e(\mathbf{x}) = \Pr(Z = 1 \mid \mathbf{x})$. See Sect. 9.2
(r_T, r_C)	For any one person, r_T is the response the person would exhibit under treatment, and r_C is the response this same person would exhibit under control. See Sect. 2.2.1
$(\mathbf{r}_T, \mathbf{r}_C)$	Vectors $\mathbf{r}_T = (r_{T11}, r_{T12}, \ldots, r_{TI2})^T$ and $\mathbf{r}_C = (r_{C11}, r_{C12}, \ldots, r_{CI2})^T$ of potential responses for all $2I$ subjects in I pairs. See Sect. 2.2.1
R	Observed response, that is, $R = r_T$ if $Z = 1$ and $R = r_C$ if $Z = 0$, so $R = Zr_T + (1 - Z)r_C$. See Sect. 2.1.2
$\mathbf{R}$	Observed responses for the $2I$ subjects in the I matched pairs, $\mathbf{R} = (R_{11}, R_{12}, \ldots, R_{I2})^T$. See Sect. 2.1.2
sgn	$\text{sgn}(a) = 1$ if $a > 0$ and $\text{sgn}(a) = 0$ otherwise. Not to be confused with its sibling sign. See Sect. 2.3.3

© Springer Nature Switzerland AG 2020

P. R. Rosenbaum, *Design of Observational Studies*, Springer Series in Statistics,
https://doi.org/10.1007/978-3-030-46405-9

sign	$\text{sign}(a) = 1$ if $a > 0$, $\text{sign}(a) = 0$ if $a = 0$, $\text{sign}(a) = -1$ if $a < 0$. Not to be confused with its sibling sgn. See Sect. 2.8
s_i	$s_i = 1$ if $\lvert Y_i \rvert > 0$ and $s_i = 0$ if $\lvert Y_i \rvert = 0$. See Sect. 2.3.3
u	An unobserved covariate. See Sect. 2.1.2
$\mathbf{u}$	Unobserved covariates for the $2I$ subjects in the I matched pairs, $\mathbf{u} = (u_{11}, u_{12}, \ldots, u_{I2})^T$. See Sect. 2.1.2
Y_i	Treated-minus-control difference in observed responses within pair i, that is, $Y_i = (Z_{i1} - Z_{i2})(R_{i1} - R_{i2})$. See Sect. 2.1.2
$\mathbf{Y}$	Treated-minus-control differences in observed responses for I pairs, $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_I)^T$. See Sect. 2.1.2
$\mathscr{F}$	The values of $(r_T, r_C, \mathbf{x}, u)$ for all subjects. Notice that these quantities do not change when Z changes, whereas R does change when Z changes. In this specific sense, $\mathscr{F}$ is fixed in Fisher's theory of randomization inference. See Sect. 2.2.1. In simple experiments, the dose of treatment is the same for every treated subject, and so doses are not explicitly mentioned when $\mathscr{F}$ is first defined in Sect. 2.2.1. However, if there are doses d_i that vary by pair i, as in Sect. 5.2.5, or that vary by individual, (d_{Tij}, d_{Cij}), as in Sect. 5.3, then these doses are also part of $\mathscr{F}$. In a strict sense, the doses are always present in $\mathscr{F}$, but when the doses are constant, there is no need to mention them
$\mathscr{Z}$	The set containing the 2^I possible values of $\mathbf{z} = (z_{11}, z_{12}, \ldots, z_{I2})^T$ of $\mathbf{Z}$. So $\mathbf{z} \in \mathscr{Z}$ if $z_{ij} = 0$ or $z_{ij} = 1$ with $z_{i1} + z_{i2} = 1$. In conditional probabilities, conditioning on $\mathscr{Z}$ means conditioning on the event $\mathbf{Z} \in \mathscr{Z}$. See Sect. 2.2.3
d_i	The dose at which the treatment is given in pair i. See the discussion of dose–response in Sect. 5.2.5
(d_{Tij}, d_{Cij})	For person j in pair i, d_{Tij} is the dose of treatment this person would accept if assigned to treatment, $Z_{ij} = 1$, and d_{Cij} is the dose of treatment this person would accept if assigned to control, $Z_{ij} = 0$. See the discussion of instrumental variables in Sect. 5.3
D_{ij}	For person j in pair i, D_{ij} is the dose actually received, $D_{ij} = Z_{ij} d_{Tij} + (1 - Z_{ij}) d_{Cij}$. See the discussion of instrumental variables in Sect. 5.3
Γ	Sensitivity parameter in a sensitivity analysis. Two subjects with the same observed covariates may differ in their odds of treatment by at most a factor of Γ. The parameter Γ is a key element in the "second model" for an observational study in Chap. 3
$P_{\min}$ and $P_{\max}$	Interval of possible P-values in a sensitivity analysis. See Sect. 3.5, specifically Table 3.2

$\widetilde{\Gamma}$	Design sensitivity. In the favorable situation, the power of a sensitivity analysis tends to 1 as the sample size increases if $\Gamma < \widetilde{\Gamma}$ and the power tends to 0 if $\Gamma > \widetilde{\Gamma}$. See Chap. 15
(Λ, Δ)	The one parameter Γ may be given a two-parameter interpretation via the amplification $\Gamma = (\Lambda \times \Delta + 1)/(\Lambda + \Delta)$. Here, Λ describes the relationship between Z and u, and Δ describes the relationship between r_C and u. A single value of Γ corresponds with a curve defined by (Λ, Δ). See Sect. 3.6

Acronyms

AAA	ACE inhibitor after anthracycline randomized trial. See Sect. 5.3
AFDC	Aid to Families with Dependent Children, a federally funded public assistance program in the USA
CI	Confidence interval
FARS	Fatality Analysis Reporting System. FARS records information about fatal road accidents in the USA
FTE	Full-time equivalent employment
GO	Gynecologic oncologist
HL	Hodges–Lehmann point estimate
IC	Interstitial cystitis, a urologic disorder
iid	Independent and identically distributed. See Sect. 15.1
IV	Instrumental variable
ICDB	Interstitial Cystitis Data Base
Medicare	U.S. government program providing health care to people over the age of 65
MN	Micronuclei, a common measure in genetic toxicology
MO	Medical oncologist
NHANES	US National Health and Nutrition Examination Survey
NICU	Neonatal intensive care unit
NJ	New Jersey, a state in the USA
NSAID	Nonsteroidal anti-inflammatory drug
NSW	National Supported Work Demonstration, a randomized experiment evaluating a program intended to aid the transition to employment
PA	Pennsylvania, a state in the USA
PMA	Postmenstrual age, a proxy for age from conception
`R`	A statistical software package. http://cran.r-project.org/ `R` is distributed free
SAS	A statistical software package. http://www.sas.com/

© Springer Nature Switzerland AG 2020

P. R. Rosenbaum, *Design of Observational Studies*, Springer Series in Statistics,
https://doi.org/10.1007/978-3-030-46405-9

SEER	Surveillance, Epidemiology, and End Results program of the U.S. National Cancer Institute. http://seer.cancer.gov/
sib-TDT	Sibship transmission disequilibrium test
stata	A statistical software package. http://www.stata.com/
TDT	Transmission disequilibrium test
TOST	Two one-sided tests procedure for equivalence testing

Glossary of Statistical Terms

Bonferroni Inequality If A and B are two events, then the probability of the event "A or B" is at most the sum of the probabilities of A and of B; that is, $\Pr(A \cup B) \leq \Pr(A) + \Pr(B)$. The inequality applies also with more than two events. The Bonferroni inequality is used in many different ways, but it shows up in this book in connection with testing more than one hypothesis. If there are two hypotheses, where A is the event of falsely rejecting the first hypothesis, and B is the event of falsely rejecting the second hypothesis, then the chance of falsely rejecting at least one hypothesis is at most $\Pr(A) + \Pr(B)$. In particular, if each test is done at level 0.025, then the chance of at least one false rejection in two tests is at most $0.025 + 0.025 = 0.05$. This use of the Bonferroni inequality is attractive because it is simple and very general; however, there are methods that are only slightly more complex and equally general but better; see Holm, S.: A simple sequentially rejective multiple test procedure, Scand J Statist **6**, 65–70 (1979).

Boxplot A boxplot is a graph that looks like a transistor. Figure 1.1 contains several boxplots. The box contains half the observations. The box has a middle line at the median and it ends at the upper and lower quartiles. One-quarter of the observations are inside the box above the median, and one-quarter of the observations are inside the box below the median. One-quarter of the observations are above the box, and one-quarter of the observations are below the box. Extreme observations appear as individual points. Boxplots were invented by John Tukey. See Cleveland, W.W.: Elements of Graphing Data, Summit, NJ: Hobart Press, pp. 139–143.

Censored If a patient was diagnosed 2 years ago and is alive today, then this patient's survival time from diagnosis is censored at 2 years. This means that we know the patient's total survival after diagnosis is at least 2 years, but we do not know the actual survival time. Typically, some survival times are censored and others are known or uncensored. Given a choice, you would rather be censored.

© Springer Nature Switzerland AG 2020

P. R. Rosenbaum, *Design of Observational Studies*, Springer Series in Statistics,
https://doi.org/10.1007/978-3-030-46405-9

Close Match A pair is closely matched for a covariate if the two individuals in the pair have very similar values of that covariate. It is generally very difficult to obtain a close match on many covariates at the same time. See also "covariate balance" for a very different notion.

Consistent Estimate A consistent estimate is one that gets it right if the sample size is large enough. An estimator $\hat{\theta}$ is a consistent estimate of a parameter θ if $\hat{\theta}$ converges in probability to θ as the sample size increases. See "convergence in probability" in this glossary.

Consistent Test A consistent test is one that gets it right if the sample size is large enough. A test of a null hypothesis H_0 is consistent against an alternative hypothesis H_A if the power of the test tends to one as the sample size increases when H_A is true.

Convergence in Distribution Convergence in distribution formalizes the notion that a random variable, say ξ_I, almost has a certain distribution. Typically in this book, ξ_I is a statistic computed from a sample of size I—for instance, a standardized version of Wilcoxon's signed rank statistic, T—and the convergence occurs as the sample size I increases. In this book, the approximating distribution is always the standard Normal distribution, $\Phi(.)$ and, to be tangible, this glossary will define convergence in distribution to the standard Normal distribution. Slightly more detail is needed to define convergence in distribution in general. A sequence of random variables, ξ_I, $I = 1, 2, \ldots$, converges in distribution to the standard Normal distribution, $\Phi(.)$, if $\lim \Pr(\xi_I \leq c) = \Phi(c)$ for all c, the limit being taken as the sample size I increases.

Convergence in Probability Convergence in probability formalizes the notion that a random variable, say ξ_I, is almost a certain constant. Typically in this book, ξ_I is a statistic computed from a sample of size I—for instance, Wilcoxon's signed rank statistic, T, divided by $I(I+1)/2$—and the convergence occurs as the sample size I increases. A sequence of random variables, ξ_I, $I = 1, 2, \ldots$, converges in probability to the real number a if $\Pr(|\xi_I - a| > \epsilon)$ tends to 0 as I increases for all $\epsilon > 0$.

Covariate A covariate is a variable measured prior to treatment assignment, and hence prior to treatment. Because it is measured prior to treatment assignment and prior to treatment, it cannot be affected by treatment. Therefore, a covariate exists in a single version, say x, because an individual has the same value of a covariate whether assigned to treatment or to control. In contrast, an outcome may be affected by the treatment, so an outcome exists in two versions, one seen under treatment, r_T, the other seen under control, r_C. See Sect. 2.1.2.

Covariate Balance If the distributions of covariates are similar in treated and control groups, then there is covariate balance. For instance, if age is one covariate, then age is balanced if the distribution of ages for treated subjects is similar to the distribution of ages for controls. Covariate balance is a property of the treated and control groups taken as two whole groups, not a property of an individual

matched pair or set. See also the entry for "close match" for a very different notion. Randomization tends to balance both observed and unobserved covariates. Matching on the propensity score tends to balance observed covariates but not unobserved covariates.

Efficiency Speaking informally, a statistical procedure is efficient if it makes the best possible use of the available sample size, producing the shortest confidence intervals, the smallest standard errors, the most powerful test, and so on. One procedure, say A, is more efficient than another procedure, say B, if A can accomplish with, say, 50 observations what B needs 52 observations to accomplish. Understandably, efficiency is a major concern in randomized controlled clinical trials, in part because each observation is expensive and time-consuming, and in part because additional observations entail treating additional patients, perhaps with a treatment that the trial itself will find to be an inferior treatment. Efficiency is not the central concern in observational studies because a limited sample size is not the principal source of uncertainty. Bias of fixed size, not the standard error, is the principal source of uncertainty in nonrandomized studies. Many observational studies would not become more compelling if the sample size were increased. In addition, in some observational studies, large samples are available from extant data.

Exclusion Restriction The exclusion restriction arises in the definition of an instrument in Sect. 5.3. Randomized encouragement to accept a treatment is an instrument for the treatment only if encouragement affects the outcome only to the extent that it affects acceptance of the treatment; this is the exclusion restriction.

Favorable Situation In the favorable situation, an observational study is free of bias from unmeasured covariates, but the investigator does not know this, and there is actually a treatment effect, but the investigator does not know this either. The investigator sees that the treated-minus-control matched pair differences in outcomes, Y_i, tend to be positive, but does not know that they tend to be positive because of a treatment effect, not because of some uncontrolled bias. It is in this favorable situation that the investigator would like to report that the treatment appears to have an effect and that appearance is highly insensitive to unmeasured biases. See Sect. 15.2.

Full Matching In a full matching, a matched set may contain one treated subject and one or more controls, or one control and one or more treated subjects. It is the "optimal form" of stratification in that it makes subjects in the same stratum as similar as possible. See Sect. 9.6.

Instrument, Instrumental Variable, IV An instrument (or instrumental variable or IV) is a randomized nudge or encouragement to accept the treatment that affects the outcome only to the extent that it affects acceptance of the treatment. See Sect. 5.3.

Level α Test A statistical test has level α if, in the worst case, it rejects a true hypothesis with probability at most α, that is, if the supremum over true hypotheses of the probability of rejection is less than or equal to α. Compare the level of a test

with the size α of a test as defined later in the glossary. In typical use, the level of the test is a promise about the test's performance and the size is a fact about its performance, where the achieved fact may be better than the promised performance. Suppose that a randomization test, such as Wilcoxon's signed rank test, is applied in a paired randomized experiment, as in Chap. 2, and the null hypothesis is rejected if the P-value is less than or equal to 0.05. Then the test will have level $\alpha = 0.05$. However, if $|\mathscr{Z}|$ is not divisible by 20, then the size of the test will be a little smaller than 0.05. In this case, one is advertising false rejections of true hypotheses at a 5% rate, but the rate is a tad better, a bit lower, than 5%. See E.L. Lehmann's book Testing Statistical Hypotheses, New York: John Wiley (1959).

Logit Model, Logistic Regression If p is the probability of an event, then $p/(1-p)$ is the odds of the event, and $\log(p/(1-p))$ is the logit or log-odds. If you were to build a linear model for p in terms of predictors, then the model might produce fitted probabilities below 0 or above 1. If the logit is linear in predictors, the probabilities stay between 0 and 1. Linear logit or logistic models assert that logits of probabilities are linear in predictors. The models are typically fitted by the method of maximum likelihood. In this book, logit models are used to estimate propensity scores, but propensity scores can be estimated in other ways, and logit models have other uses. Logit models have many attractive technical properties; see D. Cox and E. Snell, Analysis of Binary Data, New York: Chapman and Hall/CRC (1989).

Lowess Locally weighted scatterplot smoother. Lowess draws a curve through the points of a scatter plot. Figure 8.2 displays a lowess smooth. The curve is estimated locally, in the sense that, at a given value of x, the closest points (x,y) to x are given the greatest weight. Lowess was invented by William Cleveland. See Cleveland, W.W.: Elements of Graphing Data, Summit, NJ: Hobart Press, pp. 168–180.

Marginal Distribution The distribution of one random variable is its marginal distribution. The term marginal distribution is typically used when there are several random variables in the current discussion and the speaker wishes to emphasize that the distribution is for one random variable, ignoring the others. The situation is similar to that of the fellow who was delighted to learn that he had been speaking prose all his life.

Monte Carlo By tradition, algorithms that use random numbers are named for cities where people gamble. For example, there is Monte Carlo integration and there are Las Vegas algorithms. Monte Carlo integration means finding the value of an integral using random numbers. Typically, in statistical work, the integral is the expectation of a random variable. In this case, the Monte Carlo algorithm samples many independent copies of the random variable and takes their average. People sometimes use the terms Monte Carlo and simulation as if they meant the same thing, but that is not correct. Monte Carlo evaluates integrals, whereas simulation may be used for various purposes. Monte Carlo is a substitute for numerical integration, and as such it should be carried through to be accurate to several significant figures, so the value of the integral is known for practical purposes. When

someone says that such and such can be determined by Monte Carlo, the implication is that it may not be worth the time to go into further mathematical details because a child could do it. Perhaps a child with a network of supercomputers.

Ostensibly Comparable "Ostensibly comparable" means comparable in terms of measured covariates, x, leaving open the question of whether people who look comparable in terms of x are actually comparable in terms of unmeasured covariates u. To say that treated and control groups are ostensibly comparable after matching is to say that matching has done its job, has controlled for x, without presuming that this matching suffices for causal inference; it is to say that step one is complete and step two is pending.

Outcome An outcome is a variable measured after treatment. Because it is measured after treatment, it can be affected by treatment. Therefore, an outcome exists in two versions, one seen under treatment, r_T, the other seen under control, r_C. In contrast, a covariate measured prior to treatment assignment exists in a single version, say x, because an individual has the same value of a covariate whether assigned to treatment or to control. See Sect. 2.2.1.

Power (of a Test) The power of a test is the probability that the test will reject the null hypothesis when the null hypothesis is false. You would like the power to be high. If the test rejects when the P-value is less than 0.05, then the power of the test is the chance that the P-value is less than 0.05 when the null hypothesis is false. The power depends upon many things. In particular, the power depends upon what is true. If the null hypothesis is false, then something else is true. For instance, consider the power of a test of the null hypothesis that the treatment has no effect, where the test rejects when the P-value is less than 0.05. If the actual treatment effect is very small, the power may be only 0.06, but if the actual treatment effect is very large, the power may be 0.99. In other words, there may be only a 6% chance of rejecting the null hypothesis if the treatment effect is very small, but a 99% chance of rejection if the effect is very large. (Remember, if you test a true null hypothesis and reject when the P-value is less than 0.05, then you typically run a 5% chance of rejecting the true null hypothesis; so, 6% power is no reason for pride.) The power also depends upon the sample size. Typically, the power is larger when the sample size is larger.

Propensity Score The propensity score is the conditional probability of assignment to treatment given the observed covariates. See Sect. 3.3.

Quantile-Quantile Plot, or qq-Plot A plot of the quantiles of one distribution against the quantiles of another. Figure 10.1 includes a quantile-quantile plot. For instance, the lower quartile of one distribution is plotted against the lower quartile of the other, and similarly for the medians, upper quartiles, and other quantiles. Quantile-quantile plots, or qq-plots, were invented by Martin Wilk and Ram Gnanadesikan. See Cleveland, W.W.: Elements of Graphing Data, Summit, NJ: Hobart Press, pp. 143–149.

Sensitivity Analysis In an observational study, the naïve model says that people who look comparable are comparable; that is, people who look the same in terms of observed covariates do not differ in consequential ways in terms of covariates that were not measured; see Sect. 3.3. A sensitivity analysis in an observational study asks how the conclusions of the study might change if people who looked comparable were actually somewhat different; see Sect. 3.4. More generally, a sensitivity analysis asks how a conclusion that depends upon assumptions might change if those assumptions are relaxed. Performing several analyses of the same data set is not a sensitivity analysis; see Note 11 of Chap. 3.

Size α Test A statistical test has size α if, in the worst case, it rejects a true hypothesis with probability α, that is, if the supremum over true hypotheses of the probability of rejection is equal to α. Compare this with a level α test. See E.L. Lehmann's book Testing Statistical Hypotheses, New York: John Wiley (1959).

Unbiased Estimate An unbiased estimate is one that is right on average. An estimator $\hat{\theta}$ is an unbiased estimate of a parameter θ if the expectation of $\hat{\theta}$ equals θ.

Unbiased Test Speaking informally, a test is unbiased if it is more likely to reject false hypotheses than to reject true hypotheses. An α level test is unbiased against a set of alternative hypotheses if the power of the test against these alternatives is at least α. See "level α test" in this glossary.

Uncensored See censored.

Further Reading

The books listed here discuss aspects of causal inference in observational studies or experiments.

References

1. Adami, H.-O., Hunter, D.J., Lagiou, P., Mucci, L.: Textbook of Cancer Epidemiology. Oxford, New York (2018)
2. Angrist, J.D., Pischke, J.S.: Mostly Harmless Econometrics. Princeton University Press, Princeton (2009)
3. Berzuini, C., Dawid, A.P., Bernardinelli, L. (eds.): Causality: Statistical Perspectives and Applications. Wiley, New York (2012)
4. Bickman, L. (ed.): Research Design. Sage, Thousand Oaks (2000)
5. Boruch, R.: Randomized Experiments for Planning and Evaluation. Sage, Thousand Oaks (1997)
6. Breslow, N.E., Day, N.E.: Statistical Methods in Cancer Research: The Analysis of Case-Control Studies. IARC, Lyon (1980)
7. Burgess, S., Thompson, S.G.: Mendelian Randomization. Chapman and Hall/CRC, Boca Raton (2015)
8. Cahuc, P., Carcillo, S., Zylberberg, A.: Labor Economics. MIT Press, Cambridge (2014)
9. Campbell, D.T., Stanley, J.C.: Experimental and Quasi-experimental Designs for Research. Rand McNally, Chicago (1963)
10. Campbell, D.T.: Methodology and Epistemology for Social Science: Selected Papers. University of Chicago Press, Chicago (1988)
11. Cox, D.R., Reid, N.: The Theory of the Design of Experiments. Chapman and Hall/CRC, New York (2000)
12. Cox, D.R., Wermuth, N.: Multivariate Dependencies: Models, Analysis and Interpretation. Chapman and Hall/CRC, New York (1996)
13. Diggle, P.J., Heagerty, P., Liang, K-Y., Zeger, S.L.: Analysis of Longitudinal Data, 2nd edn. Oxford University Press, New York (2002)
14. Evans, A.S.: Causation and Disease: A Chronological Journey. Plenum, New York (1993)
15. Evans, L.: Traffic Safety. Science Serving Society, Bloomfield Hills (2004)

© Springer Nature Switzerland AG 2020

P. R. Rosenbaum, *Design of Observational Studies*, Springer Series in Statistics,
https://doi.org/10.1007/978-3-030-46405-9

16. Fisher, R.A.: Design of Experiments. Oliver and Boyd, Edinburgh (1935)
17. Gelman, A., Hill, J.L.: Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge, New York (2007)
18. Gerber, A.S., Green, D.P.: Field Experiments. Norton, New York (2012)
19. Hernan, M.A., Robins, J.M.: Causal Inference. CRC, Boca Raton (2018)
20. Imbens, G.W., Rubin, D.B.: Causal Inference in Statistics, Social, and Biomedical Sciences. Cambridge University Press, New York (2015)
21. Khoury, M.J., Bedrosian, S., Gwinn, M., Ioannidis, J., Higgins, J., Little, J.: Human Genome Epidemiology. Oxford University Press, New York (2010)
22. Lauritzen, S.L.: Graphical Models. Oxford University Press, New York (1996)
23. Majone, G.: Evidence, Argument, and Persuasion in the Policy Process. Yale University Press, New Haven (1989)
24. Manski, C.F.: Public Policy in an Uncertain World: Analysis and Decisions. Harvard University Press, Cambridge (2013)
25. Morgan, S.L., Winship, C.: Counterfactuals and Causal Inference: Methods and Principles for Social Science, 2nd edn. Cambridge University Press, New York (2014)
26. Murnane, R.J., Willett, J.B.: Methods Matter: Improving Causal Inference in Education and Social Science Research. Oxford University Press, New York (2010)
27. Nagin, D.S.: Group-based Modeling of Development. Harvard University Press, Cambridge (2005)
28. Pearl, J.: Causality, 2nd edn. Cambridge University Press, New York (2009)
29. Reichardt, C.S.: Quasi-Experimentation. Guilford Press, New York (2019)
30. Roberstson, L.S.: Injury Epidemiology. Oxford University Press, New York (1998)
31. Rosenbaum, P.R.: Observational Studies, 2nd edn. Springer, New York (2002)
32. Rosenbaum, P.R.: Observation and Experiment: An Introduction to Causal Inference. Harvard University Press, Cambridge (2017)
33. Rothman, K.J., Greenland, S., Lash, T.L.: Modern Epidemiology, 3rd edn. Wolters Kluwer, Philadelphia (2008)
34. Rutter, M.: Identifying the Environmental Causes of Disease: How do we Decide What to Believe and When to Take Action? Academy of Medical Sciences, London (2007)
35. Salganik, M.J.: Bit by Bit: Social Research in the Digital Age. Princeton University Press, Princeton (2018)
36. Shadish, W.R., Cook, T.D., Campbell, D.T.: Experimental and Quasi-Experimental Designs for Generalized Causal Inference. Houghton-Mifflin, Boston (2002)
37. Strom, B.L., Kimmel, S.E., Hennessy, S.: Textbook of Pharmacoepidemiology, 2nd edn. Wiley, New York (2013)
38. Susser, E., Schwartz, S., Morabia, A., Bromet, E.J.: Psychiatric Epidemiology. Oxford University Press, New York (2006)
39. Susser, M.: Causal Thinking in the Health Sciences: Concepts and Strategies in Epidemiology. Oxford University Press, New York (1973)
40. Susser, M.: Epidemiology, Health and Society: Selected Papers. Oxford University Press, New York (1987)
41. van der Laan, M., Robins, J.M.: Unified Methods for Censored Longitudinal Data and Causality. Springer, New York (2003)
42. van der Laan, M.J., Rose, S.: Targeted Learning: Causal Inference for Observational and Experimental Data. Springer, New York (2011)
43. VanderWeele, T.: Explanation in Causal Inference: Methods for Mediation and Interaction. Oxford University Press, New York (2015)
44. Wainer, H. (ed.): Drawing Inferences from Self-Selected Samples. Routledge, New York (2010)
45. Weiss, N.S., Koepsell, T.D.: Epidemiologic Methods, 2nd edn. Oxford University Press, New York (2014)
46. Wild, C., Vineis, P., Garte, S.: Molecular Epidemiology of Chronic Diseases. Wiley, New York (2008)

47. Willett, W.: Nutritional Epidemiology. Oxford University Press, New York (2013)
48. Wolpin, K.I.: The Limits of Inference without Theory. MIT Press, Cambridge (2013)
49. Wooldridge, J.M.: Econometric Analysis of Cross Section and Panel Data. MIT Press, Cambridge (2010)
50. Wu, C.F.J., Hamada, M.S.: Experiments: Planning, Analysis and Optimization. Wiley, New York (2009)

Suggested Readings for a Course

The articles listed here are suggested supplementary readings for a course on the design of observational studies.

References

1. Angrist, J.D., Krueger, A.B.: Empirical strategies in labor economics. In: Ashenfelter, O., Card, D. (eds.) Handbook of Labor Economics, vol. 3, pp. 1277–1366. Elsevier, New York (1999)
2. Bross, I.D.J.: Statistical criticism. Cancer **13**, 394–400 (1961)
3. Campbell, D.T.: Factors relevant to the validity of experiments in social settings. Psychol. Bull. **54**, 297–312 (1957)
4. Cochran, W.G.: The planning of observational studies of human populations (with Discussion). J. R. Stat. Soc. A **128**, 234–265 (1965)
5. Cornfield, J., Haenszel, W., Hammond, E., Lilienfeld, A., Shimkin, M., Wynder, E.: Smoking and lung cancer: recent evidence and a discussion of some questions. J. Nat. Cancer Inst. **22**, 173–203 (1959). Reprinted with Discussion by David R. Cox, Jan P Vandenbroucke, Marcel Zwahlen and Joel P. Greenhouse in Int. J. Epidemiol. **38**, 1175–1201 (2009)
6. Cox, D.R.: Causality: some statistical aspects. J. R. Stat. Soc. A **155**, 291–301 (1992)
7. Dawid, A.P.: Conditional independence in statistical theory (with Discussion). J. R. Stat. Soc. B **41**, 1–31 (1979)
8. Fisher, R.A.: Design of Experiments, Chapter 2. Oliver and Boyd, Edinburgh (1935)
9. Frangakis, C.E., Rubin, D.B.: Principal stratification in causal inference. Biometrics **58**, 21–29 (2002)
10. Freedman, D.A.: On the so-called 'Huber sandwich estimator' and 'robust standard errors'. Am. Stat. **60**, 299–302 (2006)
11. Friedman, M.: The methodology of positive economics. In: Essays in Positive Economics. University of Chicago Press, Chicago (1953)
12. Gelman, A.: Multilevel (hierarchical) modelling: what it can and cannot do. Technometrics **48**, 432–435 (2006)
13. Greenhouse, S.W.: Jerome Cornfield's contributions to epidemiology. Biometrics, Supplement **28**, 33–45 (1982)
14. Hamilton, M.A.: Choosing the parameter for a 2×2 table or a $2 \times 2 \times 2$ table analysis. Am. J. Epidemiol. **109**, 362–375 (1979)

© Springer Nature Switzerland AG 2020

P. R. Rosenbaum, *Design of Observational Studies*, Springer Series in Statistics,
https://doi.org/10.1007/978-3-030-46405-9

15. Hansen, B.B.: Full matching in an observational study of coaching for the SAT. J. Am. Stat. Assoc. **99**, 609–618 (2004)
16. Hill, A.B.: The environment and disease: association or causation? Proc. R. Soc. Med. **58**, 295–300 (1965)
17. Imbens, G.W.: Potential outcome and directed acyclic graph approaches to causality: relevance for empirical practice in economics. National Bureau of Economic Research, Cambridge (2019), w26104
18. Meyer, B.D.: Natural and quasi-experiments in economics. J. Bus. Econ. Stat. **13**, 151–161 (1995)
19. Neyman, J.: On the application of probability theory to agricultural experiments: essay on principles, Section 9. In Polish, but reprinted in English with Discussion by T. Speed and D. B. Rubin in Statist Sci **5**, 463–480 (1923, reprinted 1990)
20. Pearl, J.: Causal diagrams for empirical research (with Discussion). Biometrika **82**, 669–710 (1995)
21. Peto, R., Pike, M., Armitage, P., Breslow, N., Cox, D., Howard, S., Mantel, N., McPherson, K., Peto, J., Smith, P.: Design and analysis of randomised clinical trials requiring prolonged observation of each patient, I. Br. J. Cancer **34**, 585–612 (1976)
22. Robins, J.M., Hernan, M.A., Brumback, B.: Marginal structural models and causal inference in epidemiology. Epidemiology **11**, 550–560 (2000)
23. Rosenzweig, M.R., Wolpin, K.I.: Natural 'natural experiments' in economics. J. Econ. Lit. **38**, 827–874 (2000)
24. Rubin, D.B.: Estimating causal effects of treatments in randomized and nonrandomized studies. J. Educ. Psychol. **66**, 688–701 (1974)
25. Rubin, D.B.: Direct and indirect causal effects via potential outcomes. Scand. J. Stat. **31**, 161–170 (2004)
26. Sobel, M.E.: An introduction to causal inference. Sociol. Methods Res. **24**, 353–379 (1996)
27. Tchetgen Tchetgen, E.J.: The control outcome calibration approach for causal inference with unobserved confounding. Am. J. Epidemiol. **179**, 633–640 (2013)
28. Vandenbroucke, J.P.: When are observational studies as credible as randomized trials? Lancet **363**, 1728–1731 (2004)
29. West, S.G., Duan, N., Pequegnat, W., Gaist, P., Des Jarlais, D.C., Holtgrave, D., Szapocznik, J., Fishbein, M., Rapkin, B., Clatts, M., Mullen, P.D.: Alternatives to the randomized controlled trial. Am. J. Public Health **98**, 1359–1366 (2008)
30. Zubizarreta, J.R., Cerda, M., Rosenbaum, P.R.: Effect of the 2010 Chilean earthquake on post-traumatic stress: reducing sensitivity to unmeasured bias through study design. Epidemiology **7**, 79–87.

Index

A
Abadie, A., 77, 103, 165, 166
abz(y), 99, 390, 409
Achinstein, P., 113, 126
Acronyms, 515
Adami, H.-O., 523
Adaptive crosscut test, **403**
Adaptive inference, 395
Additive treatment effect, *see* Treatment effect
Ahlberg, C.E., 160, 163, 170
Ahuja, R.K., 226
Aiken, L.H., 228, 314
Ainslie, G., 121, 126
Aitkin, M., 312
Alam, K., 421, 424, 443
Albers, W., 99, 103, 390, 409, 410
Aligned rank statistic
  generalization of Wilcoxon's signed rank statistic, 55
Aligned response, **55**
Almond, D., 117, 126
Almost exact matching, 234
Almqvist, C., 165, 167
Alonso, A., 7, 19, 21
Amplification, 95, 102
amplification, xii
Analytic adjustments
  Are they feasible?, 463
Anderson, I., 17, 20
Andrews, D.F., 38, 69
Angrist, J.D., 7, 20, 75, 103, 135, 137, 138, 142, 148, 150, 152, 154, 156, 157, 159, 165, 166, 291, 295, 312, 370, 379, 380, 384, 488, 523, 527
Anthony, J.C., 146, 166, 290, 295
Apel, R., 295
Approximation algorithms for matching, 450
Ares, M., 166
Armitage, P., 141, 170, 528
Armstrong, C.S., 19, 20, 102, 104, 165, 166
Armstrong, K.A., 191, 192, 199, 227, 243, 245, 258, 276, 361, 445, 449, 450, 457, 458, 467, 471
Arnold, B.F., 165, 166
Ascertainment problems, 181
Ascherio, 165, 170
Ashenfelter, O., 345, 347
Asma, S., 399, 412
Assignment algorithm, 212
Assignment problem, *see* Assignment algorithm
Athens, L., 122, 126, 470, 471
Athey, S., 20, 31, 68, 69, 71, 77, 104, 165, 166
Attenuation of unmeasured bias, 241, 453
Attributable effects, **54**
Austin, J.L., 180, 186
Austin, P.C., 223, 226
Average treatment effect, 48
Avriel, M., 226, 242

B
Bahadur efficiency, **407**
  of adaptive inference, 409
  and evidence factors, 427
  of evidence factors, 409, 442
  of an instrument, 162, 409
  of a sensitivity analysis, 333, 405, 408
Bahadur, R.R., 333, 335, 407, 410

© Springer Nature Switzerland AG 2020

P. R. Rosenbaum, *Design of Observational Studies*, Springer Series in Statistics,
https://doi.org/10.1007/978-3-030-46405-9

Baiocchi, M., 49, 156, 157, 160, 161, 163–166, 170, 256, 274, 275, 277, 299, 310, 312, 313, 457
Baker, R.M., 151, 159, 167
Bakewell-Sachs, S., 227, 277, 284, 296, 464, 472, 492–494, 501
Balanced incomplete block design, 264, 267
Banath, J.P., 418, 444
Barnard, J., 165, 166
Barnes, B.A., 489, 500
Basse, G., 31, 69
Basta, N.E., 167
Basu, S., 275, 277, 310, 312, 313
Battistin, E., 165, 167
Bauer, P., 417, 439, 443, 480, 483, 491, 492, 495, 496, 499, 500
Bazzano, L.A., 455, 457
Beaumont, J.J., 109, 116, 128, 138, 173
Beautrais, A.L., 167
Beck, C., 163, 165, 170, 275, 276, 287, 295, 296, 299, 310, 311, 313
Becker, H.S., 469, 471
Becker, S.O., 104
Bedrosian, S., 524
Behrman, J.R., 165, 167
Benjamin-Chung, J., 165, 166
Benjamini, Y., 361, 495, 500
Bennett, D.M., 69
Berger, R.L., 480, 483, 491, 492, 495, 496, 500
Bergstralh, E.J., 226
Berk-Jones theorem, 409
Berk, R.A., 131, 167
Berk, R.H., 409, 410
Bernanke, B.S., 139, 140, 167
Bernardinelli, L., 171, 336, 385, 523
Bernardini, G., 114, 126
Bertrand, M., 77, 104
Bertsekas, D.P., 212, 223, 226, 237, 242, 300, 311, 312
Berzuini, C., 171, 336, 385, 523
Besley, T., 19, 20, 75, 104
Bhattacharya, B.B., 109
Bias
  known direction, 139
  test for, 134
Bickel, P.J., 38, 69, 99, 103, 233, 242, 390, 409, 410
Bickman, L., 105, 187, 523
Bilban, M., 133, 167, 478, 479, 481, 483, 484, 486, 500
Bilimoria, K.Y., 313, 458
Biological gradient, *see* Doses
Biologic plausibility, *see* Reasons for effects
Birch, M.W., 293, 295, 402, 403, 410, 411
Bitterman, M., 134
Bittner, E., 471
Blacker, D., 13, 21
Black, S., 167
Blauw, G.J., 117, 127
Blevins, D., 296
Blokland, A.A.J., 294–297
Blouin, J.L., 102, 104, 166
Boehnke, M., 12, 20
Bold entry in index
  meaning of, **18**
Bollobás, B., 17, 20
Bonferroni inequality, 478, 489, 517
  improvements upon, 495
  versus split samples, 461
Bongiorno, D., 458
Boruch, R., 67, 69, 523
Bosk, C.L., 470, 471
Bottleneck assignment problem, 240
Bound, J., 140, 151, 159, 167
Bowers, J., 242, 243
Boxplot, 517
Braitman, L.E., 226
Brannath, W., 417, 439, 443
Branson, Z., 470, 471
Braun, J., 227, 313
Brensinger, C.M., 296, 501
Breslow, N., 141, 170, 523, 528
Brew, B.K., 165, 167
Brooks-Gunn, J., 105
Brooks, J.M., 453, 457
Bross, I.D.J., 85, 104, 186, 527
Brown, A., 117, 128
Brown, B.M., 389, 397, 411
Brown, J.R., 113, 126
Bruce, M.L., 241, 242
Bruckdorfer, K.R., 3, 21
Brumback, B., 528
Bunker, J.P., 489, 500, 501
Burgess, S., 523
Burman, C.F., 495, 500
Busch, A.B., 165, 173, 501

C

$\mathscr{C}$, **97**
Cahuc, P., 118, 126, 268, 276, 523
Cain, G.S., 463, 471
Cai, T.T., 165, 168, 169
Caliendo, M., 104
Caliper
  adjusting, 208, 505
  how to set, 208, 212, 305, 505
  optimal, 240

in R, 305
for a time-dependent covariate, 286
using a penalty function, 212
Caliper matching, *see* Matching, caliper
Campbell, D.T., 18, 20, 22, 62, 71, 104, 108, 131, 132, 134, 135, 167, 172, 187, 313, 314, 445, 457, 523, 524, 527
Carcillo, S., 268, 276, 523
Card, D., 118, 126, 159, 167, 267, 268, 275, 276, 313, 346, 347
Cardinality matching, 241, **255**
Carnegie, N.B., 102, 104
Carpaneto, G., 226
Case reports from the Massachusetts General Hospital, 468
Case, A., 19, 20, 75, 104
Case-crossover study, 345
Cattaneo, M.D., 165, 166
Caughey, D., 384, 500
Ceballos, J.M., 142, 166, 170
Censored, 517
Censored survival times
example, 195
Cerda, M., 109, 173, 359, 362, 528
Chamberlin, T.C., 113, 126
Chambers, J., 313
Characteristic function, 103
Cheng, J., 333, 336
Cheng, Y., 165, 167
Chen, Q., 418, 443, 444
Chen, W., 108, 227, 472
Chen, Z., 418, 443, 444
Chi, S.S., 102, 104
Choudhri, E.U., 139, 167
Cioran, E.M., 129, 167
Clark, A.S., 228, 445, 447, 457, 458
Clark, B.J., 172
Clark, C.R., 120, 126
Clatts, M., 22, 173, 528
Clayton, B., 412
Cleveland, W.S., 196, 199, 242
Clinton, S.K., 125, 126
Closed testing, 495, 496
and evidence factors, 427
Close match, 518
Clustered treatment assignment, 241
Cnaan, A., 7, 20, 149–152, 159, 168, 172
Cochran, W.G., 4, 18, 20, 69, 141, 167, 226, 242, 264, 276, 415, 443, 463, 471, 477, 494, 500, 527
Cohen, T., 165, 170
Coherence, 77, 134–136, 165, 363, 364, 488
discovered in data, 371–374, 489
and equivalence tests, 490
hedged bets, 373, 489
impact on design sensitivity, 364–367, 369–374
Scheffé's method, 371–374
Coherent signed rank statistic, 136, 137, 165, 364, 369–374, 380, 383
Colditz, G.A., 169, 176, 177
Coleman, J., 463, 471
Coley, N., 126
Colford, J.M., 165, 166
Collins, A.R., 418, 443
Cologne, J., 168
Comparability
checking, 5, 82, 192, 194, 198, 229, 281, 286, 287, 306
Completely randomized experiment, **421**
Complier, **157**, 159, 162
Complier average causal effect, 157
conditional independence, $\perp\!\!\!\perp$, 79
Confidence interval
for an additive constant effect, 45
for attributable effects, 60, 61
for a multiplicative effect, 52
for a Tobit effect, 53
Conley, T.G., 167
Conover, W.J., 62, 69, 351, 356, 359, 361, 411, 424, 443
Consistent estimate, 518
Consistent test, 518
Constant treatment effect, *see* Treatment effect
Control by systematic variation, *see* Systematic variation
Control construct, 138
Control groups
entwined, 446, 447
overlapping, 446
prodded, 454
tapered, 467
two, 77, 132–134, 271, 445, 450, 479–482, 487, 507
Control outcome, 138, 367
Convergence
in distribution, 518
in probability, 518
Cook, T.D., 18, 20, 22, 62, 67, 69, 71, 108, 131, 167, 172, 314, 524
Cook, W.J., 226, 242, 275, 276
Copas, J.B., 104
Corner test, 400
Cornfield, J., 19, 20, 85, 104, 527
Cosets, 431
transversal, 432
Costa, M., 201, 202, 225, 226, 246, 258
Couch, K.A., 24, 69

Coulibaly, D., 126
Counterclaims, 6, 179, 503
  anticipated, 185
  and devices, 503
  undermined, 179
Covariance adjustment, 20
Covariate balance, 192, 518
  compared to balance produced by randomization, 232
  example, 194
  in R, 306
  in randomized experiments, 201
  standardized difference, 229
  with time-dependent covariates, 281, 286
Covariates, 5, 518
  distinguished from outcomes, 29
  missing values, 238, 303
  principal unobserved, 86
  time-dependent, 281, 284, 286
Cox, D.R., 31, 37, 67–69, 141, 170, 175, 177, 264, 289, 294, 295, 385, 461, 471, 494, 500, 523, 527, 528
Cox, G.M., 264, 276
Cox, R.B., 120, 126, 127
Crama, Y., 446, 450, 457
Crosscut test, **400**, 410
  adaptive, 410
  as an evidence factor, 426
  stratified, 402
  stratified but unmatched, 403
Crump, R.K., 100, 104
Cummings, P., 131, 169, 346, 348
Cunningham, W.H., 226, 242, 276
Curtis, D., 12, 20

D
Δ, **97**
Dafoe, A., 384, 500
D'Agostino, R.B., 104
Dalgaard, P., 313
Daniel, S., 276, 445, 449, 450, 457, 458, 467, 471
Darnell, R., 312
Darwin, C., 111, 126
Dasgupta, T., 470, 472
Data
  DNA adducts from aluminum production, 360
  DNA-protein cross-links, 225, 258
  Lalonde, 68
  lead in the blood of smokers, 410
  Maimonides' rule, 166, 383
  minimum wage and employment, 275
  money for college, 312
  National Supported Work Experiment, 68
  paint as a cause of genetic damage, 166
  safety belts in vehicle crashes, 186
  smoking and periodontal disease, 383, 410, 443
  tannery workers, 443
  welding fumes and DNA damage, 102
data
  smoking and periodontal disease, 383
Data sets, 68, 102, 166, 186, 225, 258, 275, 312, 383, 410, 443
Davey Smith, G., 3, 21, 126, 127, 165, 167, 169
Dawid, A.P., 68, 69, 79, 104, 171, 336, 385, 523, 527
Dawid's notation for conditional independence, 79
Day, N.E., 523
Dehejia, R.H., 24, 68, 69, 104
De Leuuw, J., 131, 167
Dell'Amico, M., 212, 226
DelloVigna, S., 126
Delves, H.T., 412
DeMets, D.L., 67, 70
Derigs, U., 165, 167, 275, 276, 287, 295, 300, 313
Des Jarlais, D.C., 22, 173, 528
Design sensitivity, 224, **330**, 357, 363, 504, 508
  and adaptive inference, 395
  and Bahadur efficiency, 408
  and choice of test statistic, 387
  for coherence, 366
  of the crosscut test, 401
  and clustered treatment assignment, 241
  and dose-response, 374–379, 398, 401
  with an instrument, 162
  of m-statistics, 394
  and multiple controls, 19, 224
  of a new U-statistic, 391
  optimal, 366
  and split samples, 461–463
  of step-rank statistics, 389
  with an unaffected outcome, 367
  with uncommon effects, 358
  and unit heterogeneity, 343
design sensitivity
  and dose-response, 375
Devices, 132, 363, 503
Diagnostics, 229
Difference-in-differences, 11, 274, 367
Differential effect of two treatments, 145, 290, 508
Diggle, P.J., 523
Ding, P., 49, 69, 70

Diprete, T.A., 104
Directional penalty, 237
Disambiguation, 132, 494
  and elaborate theories, 477
  planning analysis, 477
Discontinuity design, 131
Disparities research, 447
Distance matrix, 205
  in R, 304
  time-dependent, 281, 286
Distribution free statistic, 43
Dmitrienko, A., 480, 495, 501
Dokouhaki, P., 500
Donner, A., 241, 242
Dorans, N.J., 361, 385, 472
Dorn, H.F., 4
Dose-response
  adaptive crosscut test, 403
  crosscut test, 400
  and design sensitivity, 398
Doses
  and design sensitivity, 374
  Are they any good?, 508
  dose-response, 141–143, 145, 363, 374–379, 400
  fixed, 142
  just two, as different as possible, 141
  measured with error, 142, 379
  nonreactive, 382
  potential, 150
  purely reactive, 382
  reactive, 382
Dose-weighted signed rank statistic, 142, 375, 380, 382
Drake, C.M., 109, 116, 128, 138, 173
Dretske, F., 121, 126
Dreyfus, J., 296
Drive for system, 124
  diet and prostate cancer, 125
  epidemiology of dementia, 125
  health insurance, 125
Duan, N., 22, 173, 528
Du, W., 68, 71
Duflo, E., 77, 104
Du, J.T., 165, 166
Dwass, M., 424, 443
Dynarski, S.M., 11, 12, 20, 77, 104, 299, 313

E
Ebrahim, S., 3, 21
Eckles, D., 31, 69
Eddings, W., 82, 105
Edmonds, J., 275, 276
Effect modification, 255
  and cardinality matching, 256, 257
  and multiple testing, 257
Effect ratio, **156**
Efficiency, **407**, 519
  of a sensitivity analysis, 333, 405
Efron, B., 38, 69
Eguchi, S., 104
Eichengreen, B., 139, 167
Eickhoff, I., 109, 336
Elaborate theories, 477, 494
  advocated by R.A. Fisher, 477
  ambition not penalized, 496
  coherence, 477, 488
  and disambiguation, 477
  equivalence and difference, 492
  planning analyses to evaluate, 477, 478, 495
  and sensitivity analysis, 496
  with two control groups, 477, 487
Encouragement design, 148–150
Entwined comparison groups, 447
Equivalence test, 490, 508
Erasmus
  Nota res mala optima, 464
Ercumen, A., 165, 166
Ertefaie, A., 157, 159, 164, 165, 168, 274–276, 333, 335, 409, 411, 470, 471
Escobar, G.E., 227, 277, 284, 296, 464, 472, 492–494, 501
Estroff, S.E., 470, 471
Evans, A.S., 523
Evans, L., 10, 20, 146, 168, 181, 186, 523
Evans, W.N., 291, 295
Even-Shoshan, O., 108, 191, 192, 199, 227, 228, 243, 245, 258, 277, 284, 296, 314, 361, 445, 447, 457, 458, 464, 472, 492–494, 501
Evidence factors, 275, 405, **415**
  Bahadur efficiency, 409
  and replication, 415
  time/geography, 442
  within/between institutions, 442
Ewens, W.J., 12, 22
$e(\mathbf{x})$, 80
Exact matching, 192
  compared to fine balance, 252
Examples
  AAA randomized trial, 149–151
  advertising and prices, 120
  Alzheimer disease, 13
  Alzheimer disease and NSAIDs, 146
  ambient light and auto accidents, 130
  apolipoprotein E, 12

Examples (*cont.*)
cesarian delivery of premature babies, 158
data sets, 410, 443
diet and prostate cancer, 125
disability insurance, 140
DNA adducts from aluminum production, 353, 360
DNA-protein cross-links, 201, 246
DNA strand breaks and welding, 89
Dutch famine, 117, 118
Furman v. Georgia, 130, 131
Galileo's experiments, 112
gangs and violence, 287
gold standard, 139
growing up in a poor neighborhood, 130, 131
gun buybacks, 121, 122
handguns, 116, 138, 140
heart transplant, 279
heroin or cocaine addiction, 119
interstitial cystitis, 280, 490
joint taxation, 130, 131
lead in the blood of smokers, 393, 397, 410
lead mining, 133, 478, 479
Maimonides' rule, 7, 8, 75, 135–137, 142, 150–154, 166, 366, 370, 374, 379, 380
maturity at discharge from the NICU, 284, 464–467, 492–494
minimum wage, 118, 267, 346
money for college, 10–12, 299
motorcycle helmets, 346
National Supported Work Demonstration, 24
National Supported Work Experiment, 349
nature's natural experiment, 12
neonatal intensive care units, 160, 163
outcomes of newly trained surgeons, 445
ovarian cancer, 191, 229, 245, 349, 392, 397
paint as a cause of genetic damage, 142, 166, 374
permanent income hypothesis, 114
posttraumatic stress, 359
public versus private high schools, 463
regional or general anesthesia, 161
road hazards, 345
safety belts and injuries, 181
seat belts, 9, 10, 146
sleep and auto accidents, 131
smoking and periodontal disease, 399, 410, 416, 426–428, 443
tannery workers, 418, 443
twins and workforce participation, 291
twins with differing education, 345
vehicle mass and safety, 292
vitamins, 3
Excess travel time, 160, 161, 163
Exclusion criteria, 6, 280
Exclusion restriction, 148, 152, 156, 158, 519
and complier average causal effect, 157
Exclusive hypotheses, 482, 493, **497**
for equivalence tests, 483, 490
Exiting the treatment group, 7
Experimental economics, 123
Exterior match, 446, 447, 467
External validity, *see* Validity

F
$\mathscr{F}$, **29**
fixed or random, 96
Factorial design, 145
Fast Fourier transform, 103
Fatal Accident Reporting System (FARS), 10
Favorable situation, 357, 519
for coherence, 365
defined, **326**
for unit heterogeneity, 339
Feldman, H.I., 106
Feller, A., 31, 69
Feller, W., 258
Fenech, M., 104, 168
Fergusson, D.M., 167
Fermi, L., 114, 126
Feyerabend, P., 113, 126
Fine balance, 194, 245
compared to exact matching, 234, 252
compared to propensity scores, 245
exact, 246
and exact matching, 252
how to do it, 247
with multiple controls, 248
near-fine balance, 253
not exact, but best possible, 249
refined, 254
strength K, 255
Fine, J.P., 106
Fishbein, M., 22, 173, 528
Fisher, M.L., 237, 242
Fisher, R.A., 18, 20, 38, 67, 69, 181, 186, 477, 494, 495, 524, 527
and Mill's method of difference, 337, 344
randomization as the reasoned basis for inference, 37
sharp null hypothesis, 28, 421
Fisher's exact test, 202, 231
Flannagan, M.J., 130, 172

Fleisher, L.A., 228, 314
Fleiss, J.L., 226
Fleming, T.R., 195, 199
Flory, J.H., 168
Focused hypothesis, 136
Fodor, J., 111, 126
Fogarty, C.B., 68, 69, 100, 102, 104, 105, 168, 489, 495, 500, 508
Foster, E.M., 105
Fox, K.R., 445, 447, 457, 458
Francis, B., 312
Frangakis, C.E., 69, 105, 165, 166, 168, 169, 527
Franklin, A., 119, 127
Franklin, J.M., 82, 105, 109
Freedman, D.A., 68, 69, 527
French, B., 168, 409, 411, 427, 442, 443
Friedman, L.M., 67, 70
Friedman, M., 114, 127, 139, 168, 527
Frye, T., 168
Full matching, 218, 519
Fullmatch, 218, 223, 305
Furberg, C.D., 67, 70

G
Γ, **86**
$\widetilde{\Gamma}$, **330**
Gabriel, K.R., 443, 480, 482, 495, 496, 500
Gaieski, D.F., 100, 104, 508
Gail, M., 123, 128, 279, 280, 294, 295, 494, 500
Gaist, P., 22, 173, 528
Galileo, 112–114, 127
Gangl, M., 102, 104, 168
Gansky, S.A., 480, 495, 496, 500
Garaj-Vrhovac, V., 443
Garcia, G., 142, 166, 170
Gardeazabal, J., 166
Garfinkel, H., 471
Garfinkel, R.S., 243
Garte, S., 524
Gastwirth, J.L., 17, 20, 92, 105, 109, 389, 411
Gayer, T., 18, 20
Geertz, C., 468, 471
Gelman, A., 524, 527
Generalized linear models, 302
Generating function, 103
Generic bias, 147, 290
Gerber, A.S., 67, 70, 524
Gerring, J., 470–472
Ghosh, M., 356, 361
Giantonio, B.J., 228, 445, 447, 457, 458
Gibbons, J.D., 58, 70
Gibbons, R.D., 242, 243, 294, 296
Gibb, S.J., 167
Gilbert, P.B., 106, 108
Giovannucci, E., 125–127
Glazer, A., 120, 127
Glover, F., 240, 243
Glynn, R.J., 82, 105
Goeman, J.J., 480, 495, 500
Goetghebeur, E., 165, 168
Goldberger, A.S., 383, 385, 463, 471
Gong, T., 165, 167
Gonsebatt, M.E., 170
Gormley, T.A., 168
Gow, I.D., 165, 168
Goyal, N.K., 165, 173, 274, 275, 277
Green, D.P., 67, 70, 524
Greenhouse, J.P., 527
Greenhouse, S.W., 527
Green, K.M., 223, 225, 228
Greenland, S., 105, 345, 347, 524
Greenstone, M., 18, 20
Greevy, R., 7, 20, 70, 105, 149–152, 159, 163, 165, 168, 170, 274–276, 287, 295, 296, 299, 310, 311, 313
Grieve, R., 169
Grodstein, F., 7, 19, 21
Groeneboom, P., 411
Groeneveld, P.W., 163, 165, 166, 274, 275, 312
Groeneveld, R.A., 389, 411
Groenwold, R.H.H., 445, 453, 458
Gross, D.B., 105, 115, 127
Group
  coset, 431
  of permutation matrices, 430
  subgroup, 431
  transversal, 432
Gruber, J., 125, 127
Gu, X.S., 226
Guilbaud, O., 495, 500
Guo, Z., 168
Gwinn, M., 524

H
Haenszel, W., 20, 104, 293, 296, 402, 403, 410, 411, 527
Hahn, J., 103, 105, 131, 168
Halloran, M.E., 31, 70, 167, 333, 336
Hamada, M.S., 525
Hämäläinen, K., 68, 70
Hamermesh, D.S., 19, 20, 105, 165, 168
Hamilton, M.A., 527
Hammond, E., 20, 104, 347, 527
Hampel, F., 38, 69

Hansen, B.B., 105, 212, 218, 223, 225, 226, 241–243, 256, 258, 305, 310, 311, 313, 411, 528
Hansen, C.B., 167
Han, W.J., 105
Harada, M., 102, 104
Hargarten, S.W., 122, 127
Harre, R., 113, 127
Harris, S., 169
Haviland, A.M., 100, 105, 223, 226, 287, 295, 296, 508
Heagerty, P., 523
Heckman, J., 21, 68, 70, 168
Heidary, A., 500
Heitjan, D.F., 106
He, J., 418, 443, 444, 455, 457
Heller, R., 225, 257, 258, 333, 335, 347, 364, 366, 383, 385, 396, 411, 461–463, 470, 471, 489, 500, 508
Hennessy, J., 470, 472
Hennessy, S., 168, 524
Hernandez, E., 166
Hernán, M.A., 7, 19, 21, 524, 528
Heterogeneity, 255, 337
  controlled versus uncontrolled, 338
  and power of a sensitivity analysis, 342
  in a randomized experiment, 338
  and prognostic scores, 256
  and sensitivity to unmeasured biases, 340
  and design sensitivity, 343
  different kinds, 338
  examples of efforts to reduce heterogeneity, 344
  and Mill's method of difference, 337
  and sensitivity of point estimates, 343
Heterogeneity and causality, 337
Heterogeniety
  examples of efforts to reduce heterogeneity, 256
  reduced by matching, 256, 347
Hettmansperger, T.P., 389, 411
Hewer, A., 353, 360, 361
Hill, A.B., 19, 21, 105, 132, 134, 135, 138, 141, 168, 179, 186, 398, 528
Hill, A.S., 313, 458
Hill, J.L., 102, 104, 105, 165, 166, 524
Hinde, J., 312
Hinks, L., 412
Hirano, K., 105
Hochberg, Y., 495, 500
Hochman, L.L., 313, 458
Hodges, J.L., 47, 48, 52, 53, 55, 70, 335, 343, 347, 443
Ho, D.E., 105, 127, 168, 226, 313
Hodges-Lehmann point estimate, **47**
  of a Tobit effect, 53
  of an additive constant effect, 47
  of a multiplicative effect, 52
Hoeffding, W., 58, 70, 411
Hoek, H.W., 117, 128
Hoffer, T., 463, 471
Hogg, R.V., 444
Hollander, M., 443
Holland, P.W., 21, 70, 148, 168, 169, 171, 338, 347, 361, 385, 472
Holley, R., 17, 21
Hollister, R., 70
Holmes, A.P., 31, 70
Holm, S., 489, 500, 517
Holtgrave, D., 22, 173, 528
Hommel, G., 480, 495, 496, 500
Hornik, R., 109, 264, 267, 276
Horwood, L.J., 167
Hotz, V.J., 68, 70, 100, 104
Howard, S., 141, 170, 528
Howson, C., 119, 127
Hsu, J.C., 480, 483, 491, 495, 496, 500
Hsu, J.Y., 256–258, 347, 425, 442, 443
Hubbard, R.L., 120, 127
Huber, P., 38, 39, 64, 69, 70, 393, 411, 425, 443
Huber-White correction, 527
Hudgens, M.G., 31, 70, 106
Hughes, J.N., 294, 297
Hunter, D.J., 523
Hursting, S.D., 124, 128, 173, 179, 187
Huskamp, H.A., 165, 173, 501
Hwang, J.T.G., 483, 491, 496, 500

I

i, **25**
I, **25**
Iid, 49, 97
Imai, K., 105, 127, 168, 186, 226, 241–243, 267, 276, 313
Imbens, G.W., 20, 21, 31, 69, 77, 100, 104, 105, 131, 148, 153, 157, 159, 165, 166, 169, 267, 276, 524, 528
Instrument, 7, 148, 519
  Bahadur efficiency, 162, 409
  complier average causal effect, 157
  design sensitivity, 162
  effect ratio, 157
  strengthened, 159, 161, 163, 275, 409, 470
  strong, 151
  testing its validity, 158
  weak, 151, 153, 159, 162
  When are they useful?, 158

Instrumental variable, *see* Instrument, 519
Integer programming, 255, 275
Intention-to-treat analysis, 151
Interference between units, 31
  inference in the presence of, 31
Internal validity, *see* Validity
Intersection-union principle, 495
Intersection-union test, 483, 492
In 't Veld, B.A., 169
Invariant test statistic, 433
Ioannidis, J., 524
Isaacs, I.M., 431, 443
Isolation, **290**, 290
Itani, K., 313, 458
IV, 519

J
j, **25**
Jacobsen, S.L., 226
Jacobsohn, L., 276
Jaeger, D.A., 151, 159, 167
Jakopin, C.B., 133, 167, 478, 479, 481, 483, 484, 486, 500
Jemiai, Y., 108
Jiang, Y., 165, 173
Joffe, M.M., 106, 169, 176, 177, 267, 276
Johnson, B.A., 106
Jones, D.H., 409, 410
Joreskog, K.G., 383, 385
Ju, C., 82, 109
Jurečková, J., 39, 70

K
Kahn, H.S., 117, 127
Kalton, G., 276
Kang, H., 165, 168, 169, 223, 226
Karmakar, B., 165, 169, 257, 258, 275, 276, 382, 383, 385, 409, 411, 427, 442, 443, 446, 450, 457, 458, 472
Karmanov, V.G., 227, 243
Karrer,B., 31, 69
Katan, M.B., 21
Keele, L., 102, 106, 158, 165, 169, 186, 241–243, 258
Kelz, R.R., 158, 169, 228, 232, 242, 243, 246, 252, 254, 257–259, 311, 313, 314, 445, 447, 458, 495, 501
Kemper, P., 70
Kempthorne, O., 28, 68, 70
Kepler, J.D., 19, 20, 166
Khoury, M.J., 524
Kieser, M., 483, 491, 492, 496, 499
Kilcioglu, C., 257, 299, 311, 313
Kilgore, S., 463, 471
Kimmel, S.E., 106, 524
King, G., 105, 226, 241–243, 313
Kioumourtzoglou, M.A., 165, 170
Klar, N., 241, 242
Klopfer, S.O., 225, 226, 313
Klungel, O.H., 445, 458
Known effects, 77, 138, 507
Koch, G., 480, 495, 496, 500
Kochin, L.A., 139, 167
Koenker, R., 361
Koepsell, T.D., 524
Kolmogorov-Smirnov statistic
  bivariate, 233
Kopjar, N., 443
Korte, B., 212, 227, 237, 239, 243
Kosanke, J.L., 226
Kreuels, B., 223, 226
Krieger, A.M., 17, 20, 92, 105, 107, 109
Kronmal, R.A., 106
Kropf, S., 480, 495, 496, 500
Krueger, A.B., 20, 103, 118, 126, 165, 166, 267, 268, 275, 276, 313, 346, 347, 527
Kuhn, E.M., 122, 127
Kuhn, H.W., 212, 227
Kuhn, T., 113, 127
Kundo, D., 3, 21
Kurzweil, H., 432, 443

L
Λ, **96**
Lagiou, P., 523
Laird, N.M., 13, 21
LaLonde, R.J., 23, 24, 70, 313
LaLumia, S., 131, 169
Lambe, M., 131, 169
Landis, J.R., 296, 501
Landrum, M.B., 106
Langefeld, C.D., 12, 20
Langen, V., 109, 336
Larcker, D.F., 102, 104, 165, 166, 168
Larkin, G.L., 167
Larsson, H., 165, 167
Larzelere, R.E., 120, 126, 127
Lash,T.L., 524
Laska, E.M., 480, 496, 500
Latent trajectory groups, 287
Lauritzen, S.L., 524
Lavy, V., 7, 20, 75, 103, 135, 137, 138, 142, 150, 152, 154, 156, 159, 166, 312, 370, 379, 380, 384, 488

Lawlor, D.A., 3, 21, 165, 169
Lee, K., 257, 258, 347
Lee, M.J., 106
Lee, S.J., 106
Lehmacher, W., 488, 500
Lehmann alternative, 356
Lehmann, E.L., 45, 47, 48, 50, 52, 53, 55, 58, 67, 68, 70, 227, 335, 343, 347, 351, 356, 359, 361, 391, 411, 423, 443, 480, 492, 496, 500
Lemieux, T., 131, 169
Lenard, M., 241–243
Lendle, S.D., 82, 109
Levin, B., 226
Liang, K-Y., 523
Li, G., 126
Li, H.G., 104
Lilienfeld, A., 20, 104, 527
Li, X., 49, 69, 70
Li, Y.F.P., 276, 280, 288, 296, 490, 500
Liew, Z., 165, 170
Li, F., 169
Lin, D.Y., 106
Lin, P.E., 443
Lipsitch, M., 165, 170
Lipski, W., 240, 243
Little, J., 524
Liu, H-K., 483, 491, 496, 500
Loeys, T., 165, 168
Logan, B., 492, 501
Logan, R., 7, 19, 21
Logistic, 520
Logit, 205, 302, 520
  ordinal, 267
Lorch, S.L., 49, 69, 108, 156–158, 160, 161, 163–166, 170, 173, 227, 274, 275, 277, 284, 296, 312, 386, 442, 444, 457, 458, 463, 464, 470, 472, 473, 492–494, 501
Lou, J., 418, 443, 444
Lowess, 196, 198, 520
Lu, B., 70, 105, 106, 109, 163, 165, 170, 264, 267, 274–276, 287, 294–296, 299, 310, 311, 313, 446, 458
Ludwig, J., 170
Ludwig, J.M., 160, 161, 170, 228, 314, 445, 447, 457, 458
Lumey, L.H., 117, 127
Lu, X., 170, 500

M
M-test
  coherence, 383
Maclure, M., 345, 347
Magnanti, T.L., 226
Mahalanobis distance, 209, 210
  limitations of, 210
  rank-based, 210
Mahalanobis, P.C., 227
Maindonald, J., 227, 313
Majone, G., 524
Make your theories elaborate, *see* Elaborate theories
Makepeace, G.H., 106
Malmendier, U., 126
Manduchi, E., 225
Manski, C.F., 106, 120, 127, 170, 524
Manson, J.E., 7, 19, 21
Mantel-Haenszel test, 293, 308, 402, 403, 410
Mantel, N., 141, 170, 293, 296, 402, 403, 410, 411, 528
Marcus, R., 443, 480, 482, 495, 496, 500
Marcus, S.M., 106, 242, 243, 294, 296
Marden, J.I., 424, 444
Marginal distribution, 520
Marguart, J.W., 131, 170
Maritz, J.S., 39, 64, 70, 393, 411, 425, 444
Mark, S.D., 106
Markowski, E.P., 389, 411
Marolla, F., 117, 128
Masjedi, M.R., 500
Matching
  almost exact, 234
  with several groups, 264, 445, 467
  bias due to incomplete matching, 99
  caliper, 207
  cardinality, 241, **255**
  to control an interaction, 194
  directional penalty, 237
  with doses of treatment, 263
  efficiency, 225
  exact, 192, 236, 505
  exterior, 446, 447, 467
  fine balance, 245, 258, 312
  fixed ratio, 213, 217
  fixed versus variable ratio, 217, 218
  full, 218, 225
  ideal, 74
  in large databases, 240
  on many covariates, 80
  minimax, 240
  multilevel, 241
  with multiple controls, 213
  near-exact, 234
  near-fine balance, 253
  nonbipartite, 261, 446
  optimal, 211, 224

optimal caliper, 240
pair, 203
penalty, 234, 237
population before, 73
for the propensity score, 195
in R, 299
before randomization in experiments, 275
re-pairing a match, 256
refined balance, 254
risk-set, 279, 289
and seemingly innocuous confounding, 467
subset, 100, 102
tapered, 449, 467
and thick description, 467, 471
three groups, 450
with time-dependent covariates, 281, 284, 286
with two control groups, 264
to control an interaction, 194
variable controls, 213–217, 306
boxplot for, 506
without groups, 261
Matching before randomization, 275
Matsa, D.A., 168
May, J., 223, 226
Maynard, R., 70
McCaffrey, D.F., 82, 106
McClellan, M., 160, 170
McCullagh, P., 313
McCullough, M.L., 125, 127
McGinnis, R.E., 12, 22
McHugh, M.D., 228, 314
McKillip, J., 106, 138, 165, 170
McNeil, B.J., 106, 160, 170
McPherson, K., 141, 170, 489, 501, 528
Mealli, F., 170
Medoff-Cooper, B., 227, 277, 284, 296, 464, 472, 492–494, 501
Mehrotra, D.V., 108
Meisner, M.J., 480, 496, 500
M-estimate, 39, 64, 393, 425
Method of difference, *see* Mill, J.S.
Meyer, B.D., 21, 106, 165, 170, 445, 458, 528
Michels, K.B., 7, 19, 21
Micronuclei, 133
Mikkelsen, M.E., 69, 100, 102, 104, 105, 508
Mi, L., 227, 277, 284, 296, 464, 472, 492–494, 501
Milbourn, T., 168
Miller, D.L., 170
Mill, J.S., 113, 127, 175, 177, 348
method of difference, 337, 344
Millman, A., 227, 277, 284, 296, 464, 472, 492–494, 501
Milyo, J., 120, 127, 131, 170
Ming, K., 217, 218, 224, 227, 313, 508
Minimum cost flow in a network, 239, 249
Miratrix, L.W., 49, 69, 470–472
Missing data, 238
Mitnik, O.A., 100, 104
Mitra, N., 106
Mitrou, P.N., 127
Model-based adjustment, 20
Modularity
as an aid to transparency, 176
Mohammadi, F., 500
Monte Carlo, 520
Morgan, J.W., 169
Morgan, S.L., 524
Morral, A.R., 82, 106
Mosher, R.E., 108, 227, 472
Mosteller, F., 38, 489, 500
M-statistic, xii, 102, 393, 410
Mucci, L., 523
Mullainathan, S., 77, 104
Mullen, P.D., 22, 173, 528
Multilevel comparisons, 241, 527
Multiple control groups, *see* Control groups, two
Multiple controls, 213
Multiple operationalism, *see* Coherence
Munafo, M.R., 126, 127
Muntner, P., 455, 457
Murnane, R.J., 524
Musci, R.J., 21

N

Nagin, D.S., 100, 105, 106, 223, 226, 287, 294–297, 508, 524
Naïve model a
if naïve, why is it important?, 84
Naïve model for treatment assignment, *see* Treatment assignment
Nall, C., 241, 243
Natural experiment, 5, **75**, 77, 84, 158
Nature's natural experiment, 12
Navarro-Lozano, S., 168
Near-exact matching, 234
Near-far matching, 163
Near-fine balance, 253
Nelder, J.A., 313
Network, 239, 249
for fine balance, 249
Neuman, M.D., 160, 161, 170, 442, 444
Neumark, D., 119, 127
Newey, W.K., 106
Newhouse, J.P., 160, 170

Neyman, J., 28, 68, 70
Nichols, T.E., 31, 70
Nie, C.L., 122, 127
Nieuwbeerta, P., 294–297
Nikitin, I., 411
Niknam, B.A., 228, 313, 314, 445, 447, 457, 458
Noether, G., 405
Noether, G.E., 369, 385, 397, 411
Nonbipartite matching, 261, 275, 446
  discarding some subjects, 266
  pairing an odd number of subjects, 265, 270
  and strengthened instruments, 275
  strengthening an instrument, 163
  in risk-set matching, 286
Noncompliance, 7, 149, 150
Nonreactive dose, 382
Normand, S-L., 106, 165, 173, 242, 243, 294, 296, 501
Norvell, D.C., 346, 348
Nota res mala optima, 464
Nozick, R., 113, 127
Null distribution, 33

O
O'Brien, M.E., 122, 127
O'Brien, P.C., 195, 199
Ohsfeldt, R.I., 453, 457
Olive, P.L., 418, 444
Omstead, P.S., 400
Oosterhoff, J., 411
Opportunity, 129, 503
Optimal matching, 211
Optmatch, 218, 223, 225, 305
Ordinal logit model, 267
O'Reilly, E.J., 165, 170
Oreopoulos, P., 131, 170
Orlin, J.B., 226
Orwin, R., 276
Ostensively comparable, 476, 521
Outcomes, 5, 521
  distinguished from covariates, 29

P
$P_{\max}$, 89
Pagano, M., 70, 103, 106
Page, L.C., 241–243
Paik, M.C., 226
Papadimitriou, C.H., 227, 276
Parades, R.D., 256, 259, 347, 348
Park, S.Y., 127
Pattanayak, C., 470, 472
Pattern matching, *see* Coherence
Pearce, N., 21
Pearl, J., 524, 528
Peck, L., 165, 169
Peel, M.J., 106
Peirce, C.S., 121, 127, 415, 444
Penalties
  to correct an imbalance, 237
  directional, 237
  to implement near-exact matching, 234
  to impose a design constraint, 269
  small, 235
Penalty function, 269
  to implement a caliper, 212
Pepper, J.V., 120, 122, 127, 128
Pequegnat, W., 22, 173, 528
Perera, F.P., 382, 385
Pericak-Vance, M., 12
Peritz, E., 443, 480, 482, 495, 496, 500
Permutation matrix, 430
Peto, J., 141, 170, 528
Peto, R., 141, 170, 176, 177, 528
Petrie, C.V., 122, 128
Phillips, D.H., 353, 360, 361
π_ℓ, 73
Piantadosi, S., 67, 70
Piesse, A., 276
Pike, M., 141, 170, 528
Pimentel, S.D., 102, 106, 165, 170, 225, 232, 241–243, 246, 254, 257–259, 274–276, 299, 310, 312, 313, 445, 453–458, 467, 472, 495, 501
Pinto, D., 142, 166, 170
Pischke, J.S., 523
Pitman efficiency, 389, **407**
Planned analysis, 454, 475, 504
Platt, J., 113, 126, 127
Plausible alternatives, *see* Counterclaims
Plausible rival hypotheses, *see* Counterclaims
Pledger, G., 443
Plott, C.R., 128
Pocock, S.J., 412
Polsky, D., 163, 165, 166, 191, 192, 199, 227, 243, 245, 258, 274, 275, 312, 361
Polya, G., 119, 126, 128
Popper, K.R., 114, 126, 128, 175, 177
Posch, M., 417, 439, 443
Power, 521
  in a randomized experiment, 318
  of a sensitivity analysis, **326**
  suppositional collapse, 319–321
  of a test, **317**
  of Wilcoxon's signed rank test, 321
Pratt, J.W., 58, 70

Prentice-Wilcoxon test for censored pairs, 195
Preparata, F.P., 240, 243
Preston, D.L., 168
Primary analysis, 475, 504
Primary endpoint, 475
Principal components, 374
Principal unobserved covariate, 86
Probability generating function, 103
Prod, 454
Prognostic score, 256
Propensity score, 14, 195, 204, 205, 521
  and balancing observed covariates, 81, 204
  compared to fine balance, 245
  definition, **80**
  estimates work better, 82
  estimation in R, 302
  estimation of, 205
  evaluating a model for, 506
  logit model for, 205
  with ordered doses of treatment, 267
  with missing covariate values, 238
  poor predictions of treatment, 506
  why randomization is better, 82
  with several groups, 267
  with time-dependent covariates, 286, 289
  and unobserved covariates, 82
  versus exact matching, 234
Propert, K.J., 276, 280, 288, 296, 490, 500, 501
Proportional hazards model, 286, 289
Protocol, 7, 475
Psaty, B.M., 106
Pulleyblank, W.R., 226, 242, 276
Purely reactive dose, 382

Q
Quantile-quantile plot, 353, 521
Quasi-experiment, 6, 77, 84
Quine, W.V.O., 124, 128

R
R, 299
  convolve, 103
R_ℓ, 73
R_{ij}, **26**, 29
(r_{Tij}, r_{Cij}), **28**
R Development Core Team, 227, 276, 313, 458
$(r_{T\ell}, r_{C\ell})$, **73**
$\mathbf{R}$, **27**
$\mathbf{r}_C$, **28**
$\mathbf{r}_T$, **28**
Rachlin, H., 121, 128
Rampichini, C., 170
Randall, T.C., 191, 192, 199, 227, 243, 245, 258, 361
Randles, R.H., 444
Randomization
  defined, 30
  and internal validity, 62
  reasoned basis for inference in experiments, 37
Randomization inference
  using Huber's m-estimates, 64
  using the mean, 33
  using the stratified Wilcoxon's rank sum statistic, 422
  using Wilcoxon's rank sum statistic, 421
  using Wilcoxon's signed rank statistic, 39
Randomized experiment, 23–68
  example, 23
Rao, C.R., 365, 385
Rapaport-Kelz, R., 108, 227, 472
Rapkin, B., 22, 173, 528
Rauma, D., 131, 167
Reactive dose, 382
Reasons for effects, 121, 508
  example, 122, 123
  importance of, 124
  object for empirical investigation, 121
  theoretical, 124, 268
  versus direct evidence of effect, 123, 268
Refined balance, 254
Regression discontinuity design, *see* Discontinuity design
Reichardt, C.S., 524
Reid, N., 67, 69, 264, 523
Reinke, C., 252, 259
Reiss, P.C., 165, 168
Reitmeir, P., 488, 500
Relax IV, 225
Renyi, A., 421
Re-pairing a match, 256
Replication, 120, 159, 508
  of biases, 118
  and evidence factors, 415
  example, 119
  goal of, 120
Resnick, S.I., 421, 444
Restriction
  as an aid to transparency, 176
Rettore, E., 165, 167
Reynolds, K.D., 106, 135, 165, 170
Richardson, A., 106
Ridder, G., 105
Ridgeway, G., 82, 106
Rigdon, J., 70, 275, 277, 310, 312, 313
Rindskopf, D., 187

Risk set matching, 279–295
Risk-set matching, *see* Matching, risk-set
Ritov, Y., 106
Ritter, G., 296
Rival explanations, *see* Counterclaims
Rivara, F.P., 140, 173
Roberts, A.L., 165, 170
Robertson, L.S., 345, 348, 524
Robins, J.M., 7, 19, 21, 106–108, 295, 296, 524, 528
Robinson, J., 28, 68, 70
Rogers, W.H., 38, 69
Rohe, A., 275, 276
Romano, J.P., 45, 50, 68, 70
Romano, P.S., 108, 313
Romer, D., 128
Rosen, A.K., 77, 108
Rosenblat, J.D., 361
Rosenthal, R., 167, 457
Rosenzweig, M.R., 19, 22, 108, 165, 172, 528
Rose, S., 524
Rosnow, R.L., 167, 457
Rossi, P.E., 167
Ross, R.N., 191, 192, 199, 227, 228, 243, 245, 258, 361, 445, 447, 457, 458
Rothman, K., 398
Rothman, K.J., 141, 172, 524
Rotnitzky, A., 107, 108
Rouse, C., 345, 347
Roychoudhuri, R., 172
Rozelle, R.M., 187
R packages
  `approxmatch`, 453, 457
  `bigmatch`, 242, 257, 312
  `designmatch`, 257, 275, 312
  `DOS`, xii
  `rcbsubset`, 102
  `rcsubset`, 312
  `sensitivityfull`, 68, 102
  `sensitivitymult`, 68, 102, 383, 410
  `sensitivitymv`, 68, 410
  `sensitivitymw`, 68, 410
  `sensitivity2x2xk`, 360, 410
  `senstrat`, 102, 410
  `stats`, 68
  `thickmatch`, 471
  `DiPs`, 225, 242, 257, 312
  `exteriorMatch`, 448, 457
  `matchMulti`, 242
  `nbpMatching`, 165, 275, 311
  `nearfar`, 275, 312
  `optmatch`, 225, 311
  `rcbalance`, 225, 257, 312
  `DOS2`, xii, 68, 102, 166, 360, 383, 410
Ruberg, S.J., 483, 491, 496, 500
Rubin, D.B., 18, 22, 28, 68, 69, 71, 79, 80, 105, 107, 108, 148, 157, 165, 166, 168, 169, 172, 199, 203, 204, 224, 227, 238, 242, 243, 258, 314, 445, 458, 524, 527, 528
Rudolph, K.E., 102, 108
Rutter, M., 19, 22, 108, 165, 172, 524
Ryle, G., 468, 472

S

s_i, 42
Sacerdote, B.I., 169
Sachs, J., 139, 167
Saenger, G., 117, 128
Sakata, R., 168
Salganik, M.J., 524
Salsburg, D.S., 62, 69, 71, 348, 350, 354, 359, 361, 411, 424, 443
Samuelson, P.A., 111, 128
Sani, A., 445, 453, 458
Sarkar, P., 470, 472
Saunders, A.M., 12, 22
Savage, I.R., 17, 22
Savitz, D.A., 382, 385
Scharfstein, D., 107
Schatzkin, A., 123, 128
Scheffé, H., 371–373, 383, 385
Schmechel, D., 12, 22
Schneeweiss, S., 82, 105, 109
Schoket, B., 353, 360, 361
Schrijver, A., 226, 242, 276
Schuirmann, D.L., 483, 491, 496, 501
Schwartz, A.J., 139, 168
Schwartz, S., 191, 192, 199, 227, 243, 245, 258, 361
Seawright, J., 384, 470, 472, 500
Second control group, *see* Control groups, two
Seemingly innocuous confounding, 463–467
Sekhon, J.S., 19, 22, 172, 227, 470, 472
Sellers, M.M., 313, 458
Sennett, R., 130, 172
Sen, P.K., 39, 70
Sensitivity analysis, 14, 78, **84**, 84, 504, 522
  amplification, 95
  better than: association does not imply causation, 86
  with a binary outcome, 507
  with a censored outcome, 507
  for coherence, 489
  for a confidence interval, 93
  with covariance adjustment, 20
  for differential effects, 148

how to do it, 92
for an instrument, 153, 155
interpretation of Γ, 507
multiple controls, 19
example, 88
for an equivalence test, 490
for testing in order, 499
for a point estimate, 94
power of, 326
stratified, 19
for superiority/equivalence test, 492
for testing in order, 499
with time-dependent covariates, 289
and unit heterogeneity, 340
Sequentially exclusive partition, 496, **498**
exclusive hypotheses, 482
for coherence of two outcomes, 488
for equivalence tests, 483
for superiority/equivalence tests, 493
with two control groups, 482, 484, 486
Serfling, R., 391
Serfling, R.J., 412
Setting up the comparisons, 14
Sgn, 369
with ties, 61
without ties, 39
Shadish, W.R., 18, 20, 22, 62, 67, 69, 71, 108, 172, 314, 524
Shaffer, J.P., 37, 71, 482, 495, 496, 501
Shanthikumar, D.M., 102, 104
Shaper, A.G., 412
Sharp null hypothesis of no effect
defined, 28
test of, 32
Sharpe, J.E., 313, 458
Shepherd, B.E., 108
Shestov, D.B., 117, 128
Shimkin, M., 20, 104, 527
Shmoys, D.B., 446, 457, 458
Shi, P., 69, 102, 105
Sib-TDT, 12
Siddique, J., 242, 243, 294, 296
Sign, 369
Signed rank statistic, 39
Sign test, 388
Silber, J.H., 7, 20, 70, 77, 95, 97, 105, 108, 149–152, 159–161, 165, 168, 170, 172, 191, 192, 194–199, 227, 228, 232, 240, 242, 243, 245, 246, 252–254, 257–259, 276, 277, 284, 296, 311, 313, 314, 347, 349–351, 361, 442, 444–450, 457, 458, 464–472, 491–495, 501, 509
Simplifying the conditions of observation, 176
Sinharay, S., 361, 385, 472
Small, D.S., 49, 68, 69, 71, 100, 102, 104, 105, 108, 109, 134, 153, 156–161, 163–166, 168–170, 172, 173, 187, 225, 226, 241, 243, 253, 256–259, 274–277, 290, 291, 293, 296, 297, 312, 333, 335, 336, 347, 348, 364, 366, 379, 382, 383, 385, 386, 396, 409, 411, 412, 425, 427, 442–446, 450, 453–458, 461–463, 467, 470–473, 489, 495, 500, 501, 508
Smith, H.L., 224, 228, 314
Smith, J.A., 68, 71
Smith, P., 141, 170, 528
Smith, V.L., 128
Smoking and homocysteine, 455
Sobel, M.E., 31, 71, 172, 528
Software
amplification, 102
coherence, 165, 383
crosscut test, 410
dilated effect, 360
directional penalties, 242, 312
exact sensitivity analysis, 102
exterior match, 448, 457
fine balance, 312
Fisher's method of combining P-values, 442
full matching, 311
large scale matching, 312
Mantel-Haenszel test, 410
matching for thick description, 471
matching three groups, 453, 457
m-statistic, 68, 410
multilevel matching, 242
nonbipartite matching, 165, 275
refined balance, 312
Scheffé's method, 383
sensitivity analysis, 360, 409
sensitivity analysis for a mean, 102
sensitivity analysis for an m-statistic, 102
sensitivity analysis for full matching, 102
Stephenson's test, 68, 360, 410
stratified comparisons, 410
strengthening an instrument, 275
subset matching, 102
truncated product of P-values, 442
unmatched sensitivity analysis, 102
U-statistic, 360, 409
Wilcoxon statistic, 68, 102
Solari, A., 480, 495, 500
Sommer, A., 172
Sonesson, C., 495, 500
Sorensen, J.R., 131, 170

Souleles, N.S., 105, 115, 127, 128
Sparen, P., 117, 128
Sparse representation, 239, 249
Spaulding, A.B., 296
Specificity, *see* Unaffected outcome
Spieksma, F.C., 446, 450, 457
Spielman, R.S., 12, 22
Split samples, 371, 396, 461–463, 504
Srinivas, S., 386, 442, 444, 463, 470, 473
Stability in the absence of treatment, 11
Staley, K.W., 113, 128
Stanley, J.C., 523
Steiglitz, K., 227, 276
Stein, Z., 117, 128
Stellmacher, B., 432, 443
Stephenson, W.R., 55, 62, 68, 71, 351, 356, 357, 359–361, 388, 391, 410, 412
Stigler, G.J., 267, 277
Stijnen, T., 495, 500
Stochastically larger, 92, 355
Stolley, P.D., 108
Stone, M., 461, 472
Stone, R., 108
Stratified crosscut test, 402
  stratified but unmatched, 403
Strittmatter, W.J., 12, 22
Strom, B.L., 524
Strong instrument, 151
Strongly ignorable treatment assignment, *see* Treatment assignment
Structure assumption, 497
Stuart, E.A., 21, 71, 102, 105, 108, 165, 172, 223–226, 228, 242, 243, 294, 296, 313, 445, 458
Study protocol, 7
Subgroup, 431
Suissa, S., 280, 296
Sullivan, J.M., 130, 172
Summers, L.H., 165, 172
Surrogate outcomes, *see* Reasons for effects
Surrogate units (e.g., laboratory animals), *see* Reasons for effects
Susser, E., 117, 127, 128, 524
Susser, M., 19, 22, 117, 119, 128, 176, 177, 416, 444, 463, 524
SUTVA, *see* Interference between units
Swindle, T.M., 120, 126, 127
Symbols, 511
Systematic variation, 134, 338
Szapocznik, J., 22, 173, 528
Szulanski, G., 345, 348

T
T, **39**
Tamhane, A., 492, 501
Tan, Z., 172
Tapered matching, 449, 467
Tarrow, S., 470, 473
Taylor, J., 68, 71
Tchetgen Tchetgen, E.J., 31, 71, 165, 170, 172, 367, 385, 528
TDT, 12
Teng, Y., 228, 445, 447, 457, 458
Ten Have, T.R., 68, 71, 106, 241–243, 294, 296
Test
  adaptive, 395, 410
  adaptive crosscut, 403
  corner, 400
  crosscut, 400, 410, 426
  equivalence, 482, 490, 508
  intersection-union, 483, 492, 495
  invariant, 433
  level of, 519
  Mantel-Haenszel, 293, 402, 403
  m-test, 425
  power of, **317**, 521
  size of, 522
  stratified crosscut, 402
  stratified Wilcoxon, 425
  superiority/equivalence, 492
  three-sided, 495
  twice, 383, 395, 396
  U-statistic, 359, 360, 391, 426
Testing hypotheses in order, 454, 498
Testing one hypothesis twice, 383, 395, 396
Thick description, 467, **469**
Thistlethwaite, D., 131, 172
Thomas, N., 108
Thomas, Y.F., 120, 127
Thompson, S.G., 523
Three-sided test, 495
Tibshirani, R.J., 68, 71
Ties, 42, 43, 60
Tilling, K., 165, 169
Time-dependent covariates, 286
  distance matrix, 281, 286
  example, 281, 284
  propensity score, 286
  unobserved, 289
Titiunik, R., 19, 22, 169, 172
Tobin, J., 53, 71
Tobit effect, *see* Treatment effect

Todd, P., 68, 71, 131, 165, 167, 168
Tollenaar, N., 297
Tomar, S.L., 399, 412
Toniolo, P., 201, 202, 225, 226, 246, 258
TOST, 483, 491
Toth, P., 212, 226
Toulis, P., 31, 69
Trajectory groups, 287
Transmission disequilibrium test, 12
Transparency, **175**, 503
Transversal, 432
Traskin, M., 100, 108
Treatment assignment
  and the propensity score, 81
  clustered, 241
  Naïve model, 78
  naïve model, **79**, 79, 80
  strongly ignorable, 79
Treatment effect, 43–62
  additive constant, **44**, 45, 68
  attributable, 51, 54
  average, 48
  complier average, 157
  differential, 145, 508
  dilated, 360
  direct evidence of, 123
  effect ratio, 156
  general, 50
  heterogeneous, 61, 379
  hypothesis of no effect, 28
  interference between units, 31
  multiplicative, 52
  proportional to dose, 152, 376
  reason for, 123
  shift in distributions, 68
  superiority/equivalence, 492
  superpopulation, 48
  Tobit, 53
  uncommon but dramatic, 61, 349–361
  unconstrained, 50
Tremblay, R.E., 100, 105, 226, 287, 296
Tritchler, D., 70, 103, 106
Trochim, W.M.K., 108, 135, 172
Trudeau, M.E., 108, 227, 472
Truncated product of P-values, 425, 427, 442
Tseng, P., 212, 225, 226, 300, 312
Tsiatis, A.A., 106
Tukey, J.W., 38, 69, 400, 465, 473, 476, 501
Two-one-sided test procedure, 483, 491
Two-stage least squares, 151, 159

U
u_ℓ, 73
u_{ij}, **27**
$\mathbf{u}$, **27**
Ugander, J., 31, 69
Unaffected outcome, 138, 367
Unbiased estimate, 522
Unbiased test, 522
Uncensored, 522
Utada, M., 168
U-statistic, 58, 360, 391, 409, 426
  $(m, \underline{m}, \overline{m})$, 391, 426
  (5,4,5), 391
  (8,6,8), 392
  (8,7,8), 392
  and Stephenson's statistic, 359
  as an evidence factor, 426
  and Stephenson's statistic, 391
  design sensitivity, 391
Uusitalo, R., 68, 70

V
Vagero, D., 117, 128
Validity
  internal versus external, 62
Van der Klaauw, W., 131, 168
Van der Laan, M., 82, 109, 295, 296, 524
Van der Pal-de Bruin, K.M., 117, 127
Van Dyk, D.A., 267, 276
Vandenbroucke, J., 3, 18, 21, 22, 108, 165, 172, 191, 199, 468, 473, 527, 528
VanderWeele, T., 108, 524
VanderWeele, T.J., 31, 71
Van Eedeen, C., 397
Van Eeden, C., 142, 172, 385, 412
Van Schelllen, M., 295
Van Zwet, W.R., 99, 103, 390, 409, 410
Varian, H.R., 172
Varirani, V.V., 446
Vazirani, V.V., 457, 458
Velayati, A.A., 500
Vincze, I., 353, 360, 361
Vineis, P., 524
Visconti, G., 258
Volpp, K.G., 77, 108
Vuori, J., 68, 70
Vupputuri, S., 455, 457
Vygen, J., 212, 227, 237, 239, 243

W
Wager, S., 68, 71
Wahba, S., 24, 68, 69, 104

Wainer, H., 524
Waiting list, 132
Waldfogel, J., 105, 120, 127, 131, 170
Wale, C.J., 412
Walker, A.M., 458
Walker, M., 412
Wang, L.S., 109
Wang, M., 445, 447, 457, 458
Wang, W., 228, 445, 447, 457, 458
Wang, X., 165, 173
Wassmer, G., 488, 500
Watson, J., 296
Weak instrument, 151, 153, 159, 162
Webster. T.F., 382, 385
Weed, D.L., 124, 128, 173, 179, 187
Weinstein, I.B., 382, 385
Weir, B.S., 425, 442, 444
Weisskopf, M.G., 165, 170, 382, 385
Weiss, L., 72
Weiss, N., 109, 138, 141, 173, 179, 187, 524
Welch, B.L., 28, 38, 62, 68, 72
Wellenius, G.A., 382, 385
Wellford, C.F., 122, 128
Werfel, U., 88, 109, 336
Wermink, H., 297
Wermuth, N., 523
Westfall, P.H., 425, 442, 444
Westlake, W.J., 483, 491, 496, 501
West, S.G., 18, 22, 106, 135, 138, 165, 170, 173, 294, 297, 528
West, W., 294, 297
Whelton, P.K., 455, 457
White, H., 170, 500
Wiens, B.L., 480, 495, 501
Wilcox, M., 13, 21
Wilcoxon, F., 38, 72, 362, 412, 444
Wilcoxon's rank sum statistic, **421**
  stratified, **422**
Wilcoxon's signed rank statistic, **39**, 195, 375
  as a U-statistic, 391
Wild, C., 524
Wild, C.P., 383, 385
Wilk, M.B., 28, 68, 72
Willett, J.B., 524
Willett, W.C., 7, 19, 21, 525
Williams, B., 124, 128
Williams, D.M., 165, 167
Williamson, D.P., 446, 457, 458
Winship, C., 524
Wintemute, G.J., 77, 109, 116, 122, 127, 128, 138, 140, 173
Winter, S.G., 345, 348
Withers, R.L., 122, 127
Wolfe, D.A., 98, 99, 109, 444
Wolpin, K.I., 19, 22, 108, 128, 165, 172, 173, 525, 528
Wolsey, L.A., 237, 243
Wooldridge, J.M., 21, 105, 314, 525
Wright, M.A., 109, 116, 128, 138, 140, 173
Wright, P.G., 345, 348
Wu, C.F.J., 525
Wulfsohn, M., 296
Wynder, E., 20, 104, 527
Wyss, R., 82, 109

X
$\mathbf{x}_\ell$, **73**
$\mathbf{x}_{ij}$, **27**
Xu, X., 163, 165, 170, 275, 276, 287, 295, 296, 299, 310, 311, 313

Y
Y_i, **27**
$\mathbf{Y}$, 27
Yakovlev, A., 168
Yamaomoto, T., 186
Yang, D., 253, 259
Yang, F., 158, 165, 173, 275, 277, 458
Yang, L., 163, 165, 166, 274, 275, 312
Yoon, F.B., 165, 173, 258, 480, 501
Yu, B., 470, 472
Yu, B.B., 109
Yu, R., 211, 225, 228, 233, 237, 240, 242, 243, 257–259, 300, 312, 314, 456, 458, 471, 509

Z
Z_ℓ, 73
Z_{ij}, **25**
$\mathbf{Z}$, **27**
$\mathscr{Z}$
  as a set, 29
  as an event, 30
Zaheer, S., 259
Zanutto, E., 71, 109, 264, 267, 276
Zaykin, D.V., 425, 442, 444
Zeger, S., 172, 523
Zellner, A., 128
Zhang, A., 165, 169
Zhang, K., 386, 442, 444, 463, 470, 473
Zhang, M., 418, 443, 444
Zhang, N.R., 165, 173
Zhang, X., 108, 227, 472
Zhao, Q., 82, 109, 158, 169, 371, 396, 412, 463, 473, 489, 501

Zhitkovich, A., 201, 202, 225, 226, 246, 258
Zhivotovsky, L.A., 425, 442, 444
Zhou, H., 313, 458
Zou, H., 418, 443, 444
Zubizarreta, J.R., 109, 158, 160, 161, 165, 169, 170, 173, 241, 243, 252, 255–259, 274, 275, 277, 290, 291, 297, 299, 311, 313, 314, 347, 348, 359, 362, 442, 444, 458, 528
Zwahlen, M., 527
Zybert, P.A., 117, 127
Zylberberg, A., 118, 126, 268, 276, 523