Texts in Statistical Science

Statistics for Epidemiology

Nicholas P. Jewell

CHAPMAN & HALL/CRC

Statistics for Epidemiology

Nicholas P. Jewell

CHAPMAN & HALL/CRC

A CRC Press Company

Boca Raton London New York Washington, D.C.

Datasets and solutions to exercises can be downloaded at http://www.crcpress.com/e_products/downloads/.

Send correspondence to Nicholas P. Jewell, Division of Biostatistics, School of Public Health, 140 Warren Hall #7360, University of California, Berkeley, CA 94720, USA. Phone: 510-642-4627, Fax: 510-643-5163, e-mail: jewell@stat.berkeley.edu

Library of Congress Cataloging-in-Publication Data

Statistics for epidemiology / by Nicholas P. Jewell.
p. cm. — (Texts in statistical science series ; 58)
Includes bibliographical references and index.
ISBN 1-58488-433-9 (alk. paper)
1. Epidemiology—Statistical methods. I. Jewell, Nicholas P., 1952- II. Texts in statistical science.

RA652.2.M3S745 2003
614.4'072'7—dc21 2003051458

International Standard Book Number 1-58488-433-9
Library of Congress Card Number 2003051458
Printed in the United States of America 2 3 4 5 6 7 8 9 0
Printed on acid-free paper

Statistics for Epidemiology

CHAPMAN & HALL/CRC
Texts in Statistical Science Series

Analysis of Failure and Survival Data
Peter J. Smith

The Analysis and Interpretation of Multivariate Data for Social Scientists
David J. Bartholomew, Fiona Steele, Irini Moustaki, and Jane Galbraith

The Analysis of Time Series — An Introduction, Sixth Edition
Chris Chatfield

Applied Bayesian Forecasting and Time Series Analysis
A. Pole, M. West and J. Harrison

Applied Nonparametric Statistical Methods, Third Edition
P. Sprent and N.C. Smeeton

Applied Statistics — Handbook of GENSTAT Analysis
E.J. Snell and H. Simpson

Applied Statistics — Principles and Examples
D.R. Cox and E.J. Snell

Bayes and Empirical Bayes Methods for Data Analysis, Second Edition
Bradley P. Carlin and Thomas A. Louis

Bayesian Data Analysis, Second Edition
Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin

Beyond ANOVA — Basics of Applied Statistics
R.G. Miller, Jr.

Computer-Aided Multivariate Analysis, Third Edition
A.A. Afifi and V.A. Clark

A Course in Categorical Data Analysis
T. Leonard

A Course in Large Sample Theory
T.S. Ferguson

Data Driven Statistical Methods
P. Sprent

Decision Analysis — A Bayesian Approach
J.Q. Smith

Elementary Applications of Probability Theory, Second Edition
H.C. Tuckwell

Elements of Simulation
B.J.T. Morgan

Epidemiology — Study Design and Data Analysis
M. Woodward

Essential Statistics, Fourth Edition
D.A.G. Rees

A First Course in Linear Model Theory
Nalini Ravishanker and Dipak K. Dey

Interpreting Data — A First Course in Statistics
A.J.B. Anderson

An Introduction to Generalized Linear Models, Second Edition
A.J. Dobson

Introduction to Multivariate Analysis
C. Chatfield and A.J. Collins

Introduction to Optimization Methods and their Applications in Statistics
B.S. Everitt

Large Sample Methods in Statistics
P.K. Sen and J. da Motta Singer

Markov Chain Monte Carlo — Stochastic Simulation for Bayesian Inference
D. Gamerman

Mathematical Statistics
K. Knight

Modeling and Analysis of Stochastic Systems
V. Kulkarni

Modelling Binary Data, Second Edition
D. Collett

Modelling Survival Data in Medical Research, Second Edition
D. Collett

Multivariate Analysis of Variance and Repeated Measures — A Practical Approach for Behavioural Scientists
D.J. Hand and C.C. Taylor

Multivariate Statistics — A Practical Approach
B. Flury and H. Riedwyl

Practical Data Analysis for Designed Experiments
B.S. Yandell

Practical Longitudinal Data Analysis
D.J. Hand and M. Crowder

Practical Statistics for Medical Research
D.G. Altman

Probability — Methods and Measurement
A. O'Hagan

Problem Solving — A Statistician's Guide, Second Edition
C. Chatfield

Randomization, Bootstrap and Monte Carlo Methods in Biology, Second Edition
B.F.J. Manly

Readings in Decision Analysis
S. French

Sampling Methodologies with Applications
Poduri S.R.S. Rao

Statistical Analysis of Reliability Data
M.J. Crowder, A.C. Kimber, T.J. Sweeting, and R.L. Smith

Statistical Methods for SPC and TQM
D. Bissell

Statistical Methods in Agriculture and Experimental Biology, Second Edition
R. Mead, R.N. Curnow, and A.M. Hasted

Statistical Process Control — Theory and Practice, Third Edition
G.B. Wetherill and D.W. Brown

Statistical Theory, Fourth Edition
B.W. Lindgren

Statistics for Accountants, Fourth Edition
S. Letchford

Statistics for Epidemiology
Nicholas P. Jewell

Statistics for Technology — A Course in Applied Statistics, Third Edition
C. Chatfield

Statistics in Engineering — A Practical Approach
A.V. Metcalfe

Statistics in Research and Development, Second Edition
R. Caulcutt

Survival Analysis Using S—Analysis of Time-to-Event Data
Mara Tableman and Jong Sung Kim

The Theory of Linear Models
B. Jørgensen

To Debra and Britta, my very soul of life

Acknowledgments

The material in this book has grown out of a graduate course in statistical methods for epidemiology that I have taught for more than 20 years in the School of Public Health at Berkeley. I wish to express my appreciation for the extraordinary students that I have met through these classes, with whom I have had the privilege of sharing and learning simultaneously. My thanks also go to Richard Brand, who first suggested my teaching this material, and to Steve Selvin, a lifelong friend and colleague, who has contributed enormously both through countless discussions and as my local S-Plus expert. The material on causal inference depended heavily on many helpful conversations with Mark van der Laan. Several colleagues, especially Alan Hubbard, Madukhar Pai, and Myfanwy Callahan, have selflessly assisted by reading parts or all of the material, diligently pointing out many errors in style or substance. I am forever grateful to Bonnie Hutchings, who prepared the earliest versions of handouts of some of this material long before a book was ever conceived of, and who has been a constant source of support throughout. I also owe a debt of gratitude to Kate Robertus for her incisive advice on writing issues throughout the text.

Finally, my enjoyment of this project was immeasurably enhanced by the love and support of my wife, Debra, and our daughter, Britta. Their presence is hidden in every page of this work, representing the true gift of life.

Contents

1 Introduction 1
- 1.1 Disease processes . . . 1
- 1.2 Statistical approaches to epidemiological data . . . 2
  - 1.2.1 Study design . . . 3
  - 1.2.2 Binary outcome data . . . 4
- 1.3 Causality . . . 5
- 1.4 Overview . . . 5
  - 1.4.1 Caution: what is not covered . . . 7
- 1.5 Comments and further reading . . . 7

2 Measures of Disease Occurrence 9
- 2.1 Prevalence and incidence . . . 9
- 2.2 Disease rates . . . 12
  - 2.2.1 The hazard function . . . 13
- 2.3 Comments and further reading . . . 15
- 2.4 Problems . . . 16

3 The Role of Probability in Observational Studies 19
- 3.1 Simple random samples . . . 20
- 3.2 Probability and the incidence proportion . . . 21
- 3.3 Inference based on an estimated probability . . . 22
- 3.4 Conditional probabilities . . . 24
  - 3.4.1 Independence of two events . . . 26
- 3.5 Example of conditional probabilities—Berkson's bias . . . 26
- 3.6 Comments and further reading . . . 28
- 3.7 Problems . . . 29

4 Measures of Disease–Exposure Association 31
- 4.1 Relative risk . . . 31
- 4.2 Odds ratio . . . 32
- 4.3 The odds ratio as an approximation to the relative risk . . . 33
- 4.4 Symmetry of roles of disease and exposure in the odds ratio . . . 34
- 4.5 Relative hazard . . . 35
- 4.6 Excess risk . . . 37
- 4.7 Attributable risk . . . 38

4.8 Comments and further reading 40
4.9 Problems 41

5 Study Designs **43**
5.1 Population-based studies 45
5.1.1 Example—mother's marital status and infant birthweight . 46
5.2 Exposure-based sampling—cohort studies 47
5.3 Disease-based sampling—case-control studies 48
5.4 Key variants of the case-control design 50
5.4.1 Risk-set sampling of controls 51
5.4.2 Case-cohort studies 53
5.5 Comments and further reading 55
5.6 Problems 56

6 Assessing Significance in a 2 × 2 Table **59**
6.1 Population-based designs 59
6.1.1 Role of hypothesis tests and interpretation of p-values . . . 61
6.2 Cohort designs 62
6.3 Case-control designs 64
6.3.1 Comparison of the study designs 65
6.4 Comments and further reading 68
6.4.1 Alternative formulations of the χ^2 test statistic 69
6.4.2 When is the sample size too small to do a χ^2 test? 70
6.5 Problems 71

7 Estimation and Inference for Measures of Association **73**
7.1 The odds ratio 73
7.1.1 Sampling distribution of the odds ratio 74
7.1.2 Confidence interval for the odds ratio 77
7.1.3 Example—coffee drinking and pancreatic cancer 78
7.1.4 Small sample adjustments for estimators of the odds ratio . 79
7.2 The relative risk 81
7.2.1 Example—coronary heart disease in the Western Collaborative Group Study 82
7.3 The excess risk 83
7.4 The attributable risk 84
7.5 Comments and further reading 85
7.5.1 Measurement error or misclassification 86
7.6 Problems 90

8 Causal Inference and Extraneous Factors: Confounding and Interaction **93**
8.1 Causal inference 94
8.1.1 Counterfactuals 94
8.1.2 Confounding variables 99

8.1.3 Control of confounding by stratification 100
8.2 Causal graphs 102
8.2.1 Assumptions in causal graphs 105
8.2.2 Causal graph associating childhood vaccination to subsequent health condition 106
8.2.3 Using causal graphs to infer the presence of confounding 107
8.3 Controlling confounding in causal graphs 109
8.3.1 Danger: controlling for colliders 109
8.3.2 Simple rules for using a causal graph to choose the crucial confounders 111
8.4 Collapsibility over strata 112
8.5 Comments and further reading 116
8.6 Problems 119

9 Control of Extraneous Factors **123**
9.1 Summary test of association in a series of 2 × 2 tables 123
9.1.1 The Cochran–Mantel–Haenszel test 125
9.1.2 Sample size issues and a historical note 128
9.2 Summary estimates and confidence intervals for the odds ratio, adjusting for confounding factors 128
9.2.1 Woolf's method on the logarithm scale 129
9.2.2 The Mantel–Haenszel method 130
9.2.3 Example—the Western Collaborative Group Study: part 2 . 131
9.2.4 Example—coffee drinking and pancreatic cancer: part 2 . . 133
9.3 Summary estimates and confidence intervals for the relative risk, adjusting for confounding factors 134
9.3.1 Example—the Western Collaborative Group Study: part 3 . 135
9.4 Summary estimates and confidence intervals for the excess risk, adjusting for confounding factors 136
9.4.1 Example—the Western Collaborative Group Study: part 4 . 137
9.5 Further discussion of confounding 138
9.5.1 How do adjustments for confounding affect precision? . . . 138
9.5.2 An empirical approach to confounding 142
9.6 Comments and further reading 143
9.7 Problems 144

10 Interaction **147**
10.1 Multiplicative and additive interaction 148
10.1.1 Multiplicative interaction 148
10.1.2 Additive interaction 149
10.2 Interaction and counterfactuals 150
10.3 Test of consistency of association across strata 152
10.3.1 The Woolf method 153
10.3.2 Alternative tests of homogeneity 155

10.3.3 Example—the Western Collaborative Group Study: part 5 . 156
10.3.4 The power of the test for homogeneity 158
10.4 Example of extreme interaction 160
10.5 Comments and further reading 161
10.6 Problems 162

11 Exposures at Several Discrete Levels **165**
11.1 Overall test of association 165
11.2 Example—coffee drinking and pancreatic cancer: part 3 167
11.3 A test for trend in risk 167
11.3.1 Qualitatively ordered exposure variables 169
11.3.2 Goodness of fit and nonlinear trends in risk 170
11.4 Example—the Western Collaborative Group Study: part 6 171
11.5 Example—coffee drinking and pancreatic cancer: part 4 173
11.6 Adjustment for confounding, exact tests, and interaction 175
11.7 Comments and further reading 176
11.8 Problems 176

12 Regression Models Relating Exposure to Disease **179**
12.1 Some introductory regression models 181
12.1.1 The linear model 181
12.1.2 Pros and cons of the linear model 183
12.2 The log linear model 183
12.3 The probit model 184
12.4 The simple logistic regression model 186
12.4.1 Interpretation of logistic regression parameters 187
12.5 Simple examples of the models with a binary exposure 188
12.6 Multiple logistic regression model 190
12.6.1 The use of indicator variables for discrete exposures 191
12.7 Comments and further reading 196
12.8 Problems 196

13 Estimation of Logistic Regression Model Parameters **199**
13.1 The likelihood function 199
13.1.1 The likelihood function based on a logistic regression model 201
13.1.2 Properties of the log likelihood function and the maximum likelihood estimate 204
13.1.3 Null hypotheses that specify more than one regression coefficient 206
13.2 Example—the Western Collaborative Group Study: part 7 207
13.3 Logistic regression with case-control data 212
13.4 Example—coffee drinking and pancreatic cancer: part 5 215
13.5 Comments and further reading 218
13.6 Problems 219

14 Confounding and Interaction within Logistic Regression Models 221
14.1 Assessment of confounding using logistic regression models . . . 221
14.1.1 Example—the Western Collaborative Group Study: part 8 . 223
14.2 Introducing interaction into the multiple logistic regression model . . . 225
14.3 Example—coffee drinking and pancreatic cancer: part 6 . . . 227
14.4 Example—the Western Collaborative Group Study: part 9 . . . 230
14.5 Collinearity and centering variables . . . 230
14.5.1 Centering independent variables . . . 233
14.5.2 Fitting quadratic models . . . 233
14.6 Restrictions on effective use of maximum likelihood techniques . . 235
14.7 Comments and further reading . . . 236
14.7.1 Measurement error . . . 237
14.7.2 Missing data . . . 237
14.8 Problems . . . 240

15 Goodness of Fit Tests for Logistic Regression Models and Model Building 243
15.1 Choosing the scale of an exposure variable . . . 243
15.1.1 Using ordered categories to select exposure scale . . . 244
15.1.2 Alternative strategies . . . 245
15.2 Model building . . . 246
15.3 Goodness of fit . . . 250
15.3.1 The Hosmer–Lemeshow test . . . 252
15.4 Comments and further reading . . . 254
15.5 Problems . . . 255

16 Matched Studies 257
16.1 Frequency matching . . . 257
16.2 Pair matching . . . 258
16.2.1 Mantel–Haenszel techniques applied to pair-matched data . . . 262
16.2.2 Small sample adjustment for odds ratio estimator . . . 264
16.3 Example—pregnancy and spontaneous abortion in relation to coronary heart disease in women . . . 264
16.4 Confounding and interaction effects . . . 265
16.4.1 Assessing interaction effects of matching variables . . . 265
16.4.2 Possible confounding and interactive effects due to nonmatching variables . . . 266
16.5 The logistic regression model for matched data . . . 269
16.5.1 Example—pregnancy and spontaneous abortion in relation to coronary heart disease in women: part 2 . . . 271
16.6 Example—the effect of birth order on respiratory distress syndrome in twins . . . 274
16.7 Comments and further reading . . . 276

16.7.1 When can we break the match? 277
16.7.2 Final thoughts on matching 278
16.8 Problems 279

17 Alternatives and Extensions to the Logistic Regression Model 285
17.1 Flexible regression model 285
17.2 Beyond binary outcomes and independent observations 289
17.3 Introducing general risk factors into formulation of the relative hazard—the Cox model 290
17.4 Fitting the Cox regression model 293
17.5 When does time at risk confound an exposure–disease relationship? 295
17.5.1 Time-dependent exposures 296
17.5.2 Differential loss to follow-up 296
17.6 Comments and further reading 297
17.7 Problems 298

18 Epilogue: The Examples 301

References 303

Glossary of Common Terms and Abbreviations 311

Index 319

CHAPTER 1

Introduction

In this book we describe the collection and analysis of data that speak to relationships between the occurrence of diseases and various descriptive characteristics of individuals in a population. Specifically, we want to understand whether and how differences in individuals might explain patterns of disease distribution across a population. For most of the material, I focus on chronic diseases, the etiologic processes of which are only partially understood compared with those of many infectious diseases. Characteristics related to an individual's risk of disease will include (1) basic measures (such as age and sex), (2) specific risk exposures (such as smoking and alcohol consumption), and (3) behavioral descriptors (including educational or socioeconomic status, behavior indicators, and the like). Superficially, we want to shed light on the "black box" that takes "inputs"—risk factors such as exposures, behaviors, genetic descriptors—and turns them into the "output," some aspect of disease occurrence.

1.1 Disease processes

Let us begin by briefly describing a general schematic for a disease process that provides a context for many statistical issues we will cover. Figure 1.1, an adapted version of Figure 2.1 in Kleinbaum et al. (1982), illustrates a very simplistic view of the evolution of a disease in an individual.

Note the three distinct stages of the disease process: *induction, promotion*, and *expression*. The etiologic process essentially begins with the onset of the first cause of the resulting disease; for many chronic diseases, this may occur at birth or during fetal development. The end of the promotion period is often associated with a clinical diagnosis. Since we rarely observe the exact moment when a disease "begins," induction and promotion are often considered as a single phase. This period, from the start of the etiologic process until the appearance of clinical symptoms, is often called the *latency period* of the disease. Using AIDS as an example, we can define the start of the process as exposure to the infectious agent, HIV. Disease begins with the event of an individual's infection; clinical symptoms appear around the onset and diagnosis of AIDS, with the expression of the disease being represented by progression toward the outcome, often death. In this case, the induction period is thought to be extremely short in time and is essentially undetectable; promotion and expression can both take a considerable length of time.

Epidemiological study of this disease process focuses on the following questions:

- Which factors are associated with the induction, promotion, and expression of a disease? These *risk factors* are also known as *explanatory variables, predictors,*

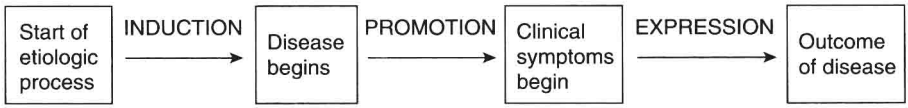

Figure 1.1 *Schematic of disease evolution.*

covariates, independent variables, and *exposure variables*. We will use such terms interchangeably as the context of our discussion changes.

- In addition, are certain factors (not necessarily the same ones) associated with the duration of the induction, promotion, and expression periods?

For example, exposure to the tubercule bacillus is known to be necessary (but not sufficient) for the induction of tuberculosis. Less is known about factors affecting promotion and expression of the disease. However, malnutrition is a risk factor associated with both these stages. As another example, consider coronary heart disease. Here, we can postulate risk factors for each of the three stages; for instance, dietary factors may be associated with induction, high blood pressure with promotion, and age and sex with expression. This example illustrates how simplistic Figure 1.1 is in that the development of coronary heart disease is a continuous process, with no obvious distinct stages. Note that factors may be associated with the outcome of a stage without affecting the duration of the stage. On the other hand, medical treatments often lengthen the duration of the expression of a chronic disease without necessarily altering the eventual outcome.

Disease intervention is, of course, an important mechanism to prevent the onset and development of diseases in populations. Note that intervention strategies may be extremely different depending on whether they are targeted to prevent induction, promotion, or expression. Most public health interventions focus on induction and promotion, whereas clinical treatment is designed to alter the expression or final stage of a disease.

1.2 Statistical approaches to epidemiological data

Rarely is individual information on disease status and possible risk factors available for an entire population. We must be content with only having such data for some fraction of our population of interest, and with using statistical tools both to elucidate the selection of individuals to study in detail (sampling) and to analyze data collected through a particular study. Issues of study design and analysis are crucial because we wish to use sample data to most effectively make applicable statements about the larger population from which a sample is drawn. Second, since accurate data collection is often expensive and time-consuming, we want to ensure that we make the best use of available resources. Analysis of sample data from epidemiological studies presents many statistical challenges since the outcome of interest—disease status—is usually binary. This book is intended to extend familiar statistical approaches for continuous outcome data—for example, population mean comparisons and regression—to the binary outcome context.

1.2.1 Study design

A wide variety of techniques can be used to generate data on the relationship between explanatory factors and a putative outcome variable. I mention briefly only three broad classes of study designs used to investigate these questions, namely, (1) *experimental studies*, (2) *quasi-experimental studies*, and (3) *observational studies*. The crucial feature of an experimental study is the investigator's ability to manipulate the factor of interest while maintaining control of other extraneous factors. Even if the latter is not possible, control of the primary risk factor allows its randomization across individual units of observation, thereby limiting the impact of uncontrolled influences on the outcome. Randomized clinical trials are a type of experimental study in which the main factor of interest, treatment type, is under the control of the investigator and is randomly assigned to patients suffering from a specific disease; other influencing factors, such as disease severity, age, and sex of the patient, are not directly controlled.

Quasi-experimental studies share some features of an experimental study but differ on the key point of randomization. Although groups may appear to differ only in their level of the risk factor of interest, these groups are not formed by random assignment of this factor. For example, comparison of accident fatality rates in states before and after the enactment of seat-belt laws provides a quasi-experimental look at related safety effects. However, the interpretation of the data is compromised to some extent by other changes that may have occurred in similar time periods (did drivers increase their highway speeds once seat belts were required?). A more subtle example involved an Austrian study of the efficacy of the PSA (prostate specific antigen) test in reducing mortality from prostate cancer; investigators determined that, within 5 years, the death rate from prostate cancer declined 42% below expected levels in the Austrian state, Tirol, the only state in the country that offered free PSA screening. Again, comparisons with other areas in the country are compromised by the possibility there are other health-related differences between different states other than the one of interest. Many *ecologic* studies share similar vulnerabilities. The absence of randomization, together with the inability to control the exposure of interest and related factors, make this kind of study less desirable for establishing a causal relationship between a risk factor and an outcome.

Finally, observational studies are fundamentally based on sampling populations with subsequent measurement of the various factors of interest. In these cases, there is not even the advantage of a naturally occurring experiment that changed risk factors in a convenient manner. Later in the book we will focus on several examples including studies of the risk of coronary heart disease where primary risk factors, including smoking, cholesterol levels, blood pressure, and pregnancy history, are neither under the control of the investigator nor usually subject to any form of quasi-experiment. Another example considers the role of coffee consumption on the incidence of pancreatic cancer, again a situation where study participants self-select their exposure categories.

In this book, we focus on the design and analysis of observational epidemiological studies. This is because, at least in human populations, it is simply not ethical to randomly assign risk factors to individuals. Although many of the analytic techniques